University of Alberta

Applications of DNA Indexing

by

Christopher J. Dambrowitz ©

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

in

Molecular Biology and Genetics

Department of Biological Sciences

Edmonton, Alberta

Spring 2002

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-68559-4

Canada

University of Alberta

Library Release Form

Name of Author: Christopher J. Dambrowitz

Title of Thesis: Applications of DNA Indexing

Degree: Doctor of Philosophy

Year this Degree Granted: 2002

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Christopher Dambrowitz

2611 NW 57th St., Unit 3
Seattle, WA USA 98107-3246

January 18, 2002

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Applications of DNA Indexing** by Christopher John Dambrowitz in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Molecular Biology and Genetics.

Kenneth L Roy

Dr. K.L. Roy

Dr. N.J. Dovichi

Dr. R.B. Hodgetts

Dr. J.F. Elliott

Dr. G. Drouin

Date January 7, 2002

Nothing in this world can take the place of persistence.

Talent will not; nothing is more common than unsuccessful men with talent.
Genius will not; unrewarded genius is almost a proverb.
Education will not; the world is full of educated failures.

Persistence and determination alone are omnipotent.

Calvin Coolidge

This thesis is dedicated to my wife, Dr. Amy Dambrowitz

and to my parents, Ivan and Cecilia Dambrowitz

and to the loving memory of Unk - John Stevenson.

ABSTRACT:

DNA indexing strategies were developed for several applications, including directed mapping and sequencing, microbial subtyping, and global gene expression studies in *Saccharomyces cerevisiae*.

The plasmid pUC19 model system was used for DNA indexing protocol and strategy development. A compound-primer indexing strategy prevented repeated-end amplification, provided the means for directional cycle sequencing of amplified indexed fragments, and eliminated primer-dimer (PD) artifact amplification in indexing PCR reactions. The pUC19 indexing system was characterized by aligning sequenced indexed templates along an indexing-based restriction map. A method for directed mapping and sequencing of prokaryotic genomes is proposed.

Protocols for indexed genomic profiling (IGP) compensated for microbial genomic digest complexity, using pooled indexer mixtures to increase information density. A database was created to characterize the predicted set of indexable restriction fragments in *Fok*I-digested *E. coli* chromosomal DNA. Genomic profile data for three *E. coli* strains were obtained using IGP protocols and compared to database predictions. Profiles were generated for several *Staphylococcus* species. IGP provides an attractive alternative to current microbial typing techniques, and may be applicable to epidemiological investigations of pathogenic *Staphylococcus* species or studies of microbial community diversity.

Modified 3'-end cDNA indexing protocols were developed for global gene expression analysis in yeast. Indexers targeting specific 3'-terminal cDNA fragments in *Fok*I-digested cDNA populations were predicted from *S. cerevisiae* sequence data. Target fragments ligated to biotinylated indexers were purified to reduce PCR complexity. Artificial poly(A)-tailed indexable constructs were used for protocol development. Distinct and reproducible 3'-end cDNA indexing profiles were generated for yeast cultures exposed to various environmental stimuli. Observations of differential expression for *GAL1* and *BOP3* were well-correlated to published data. A limited survey of gene expression changes in yeast produced results compatible with published data obtained using cDNA microarrays. However, larger 3'-end cDNA indexing data sets were poorly correlated with data from published studies of osmotic shock response in yeast. Refinement of 3'-end cDNA indexing is necessary for effective application to yeast transcriptomics. Avenues of future research were identified that may provide solutions to the unresolved challenges of cDNA indexing approaches.

Acknowledgements

I would like to express my gratitude to Dr. Norm Dovichi for his direction, support, encouragement and friendship, and for the freedom that he provides his students to chart and navigate the course of their own scientific adventures.

I would also like to express my thanks:

To Dr. Herb Heyneker of Eos Biotechnology (formerly of ProtoGene Laboratories, Inc.) for the generous gift of the NoP BamCC indexer mixes used in these investigations.

To Randy Nonay and Eric Carpenter for development of the EcoliDB and YeastORFdb software, and for crucial technical support throughout my research.

To Dr. Amy Dambrowitz for providing me with the first set of yeast RNA samples, for training a DNA guy in yeast molecular biology techniques, and for her wisdom, talent, and constant support.

And to Paul Unrau, co-inventor of DNA indexing, without whose insight and mentoring I could not have even begun.

…and who turned me into a newt.

(…I got *better*…!)

Table of Contents

1 Chapter I: Introduction to DNA Indexing 1

1.1 DNA indexing 1
1.1.1 Historical context 1
1.1.2 Principles of DNA indexing... 2
1.1.3 Estimation of system complexity for DNA indexing studies 8

1.2 Enzymes employed in DNA indexing systems...9
1.2.1 Type IIS and Type IP restriction endonucleases.. 9
1.2.2 T4 DNA ligase..12

1.3 Thesis Overview 18
1.3.1 Development of DNA Indexing Strategies for Directed Mapping and Sequencing..19
1.3.2 Bacterial Strain and Species Differentiation by Indexed Genomic Profiling (IGP)..20
1.3.3 Global Gene Expression Profiling of *Saccharomyces cerevisiae* by 3'-end cDNA Indexing...20

2 Chapter II: Development of DNA Indexing Strategies for Directed Mapping and Sequencing 23

2.1 INTRODUCTION...23
2.1.1 Selection of pUC19 as a model system for DNA indexing........................25
2.1.2 Selection of T4 DNA ligase for development of the pUC19 indexing model system...28
2.1.3 Selection of DNA polymerases for indexing applications 28

2.1.3.1 *Selection of* Taq *DNA polymerase for development of the pUC19 indexing model system* *28*

2.1.3.2 *Selection of* PfuTurbo^{TM} *DNA polymerase for amplification of indexed sequencing templates* *29*

2.2 MATERIALS AND METHODS 29

2.2.1 Evaluation and selection of DNA indexing model system 29

2.2.2 Digestion of pUC19 DNA with *Fok*I restriction endonuclease 30

2.2.3 Synthesis of indexing oligonucleotides 30

2.2.4 Annealing of indexing oligonucleotides 31

2.2.5 Ligation of indexers to pUC19 DNA: standard conditions 33

2.2.6 Amplification of indexed pUC19 fragments by Polymerase Chain Reaction (PCR) : standard conditions 34

2.2.7 Agarose gel electrophoresis of amplified indexed pUC19 fragments 35

2.2.8 UV transillumination of ethidium bromide-stained agarose gels 35

2.2.9 Protocols relating to the study of primer-dimer 36

2.2.9.1 *Isolation of primer-dimer (PD) DNA from agarose gel by syringe extraction* *36*

2.2.9.2 *Preparation of primer-dimer amplification product for cloning* ... *37*

2.2.9.3 *Blunt-end cloning of PD DNA into pUC19 sequencing vector* *38*

2.2.9.4 *Transformation of competent E. coli DH5α with vector bearing PD insert* *39*

2.2.9.5 *Amplification of insert-bearing pUC19 fragments directly from transformed* E. coli *DH5α colonies by AB PCR* *40*

2.2.9.6 *Cycle sequencing of PD insert from AB PCR product* *41*

2.2.10 Double digestion of pUC19 DNA with *Fok*I and *Sfa*NI restriction endonucleases 42

2.2.11 Amplification of indexed pUC19 fragments by *PfuTurbo*^{TM} DNA polymerase 42

2.2.12 Direct cycle sequencing of amplified indexed pUC19 fragments 43

2.2.13 Mapping and assembly of contiguous pUC19 sequences 43

2.3 RESULTS 44

2.3.1 Features of the α phosphorylated indexers and γ nonphosphorylated indexers 44

2.3.2 Digestion of pUC19 DNA by *Fok*I restriction endonuclease 47

2.3.3 Amplification of an indexed fragment following target-specific ligation of indexers 51

2.3.4 Single-primer P/P indexing of the pUC19 model system 53

2.3.5 The P/NoP indexing strategy 58

2.3.6 Establishing ligation conditions for DNA indexing 61

2.3.7 Directionality as a requirement for cycle sequencing of indexed template fragments 64

2.3.8 Generation and description of indexing PD artifact 65

2.3.9 Features of the Bam phosphorylated indexers and BamCC nonphosphorylated indexers 69

2.3.10 Use of the Bam and BamCC indexer sets for compound-primer P/NoP indexing 71

2.3.11 Compound-primer P/NoP indexing of *Fok*I-digested pUC19 fragments .75

2.3.12 Indexing of pUC19 *Fok*I fragments for sequencing template production 77

2.3.13 Amplification of gap-closing B_a and B_b sequencing templates from a *Fok*I/*Sfa*NI double restriction digest 80

2.3.14 Direct cycle sequencing of amplified indexed pUC19 fragments 83

2.3.15 Alignment of indexing-based directionally-sequenced pUC19 templates to an indexing-based restriction map of pUC19 constructed by jigsaw assembly 83

2.4 DISCUSSION 85

3 Chapter III: Bacterial Strain and Species Differentiation by Indexed Genomic Profiling (IGP) **88**

3.1 INTRODUCTION 88
  3.1.1 ABSTRACT: 90

3.2 MATERIALS AND METHODS 90
  3.2.1 Bacterial strains used in IGP investigations 90
  3.2.2 Isolation and purification of bacterial genomic DNA 91
  3.2.3 Digestion of bacterial DNA with *Fok*I restriction endonuclease 91
  3.2.4 Synthesis and annealing of indexing oligonucleotides 92
  3.2.5 Ligation of indexers to *Fok*I-digested bacterial genomic DNA: standard conditions 92
  3.2.6 Amplification of indexed bacterial genomic DNA fragments by Polymerase Chain Reaction (PCR) : standard conditions 93
  3.2.7 Agarose gel electrophoresis of amplified indexed bacterial DNA fragments on large 64-mix gels 93
  3.2.8 Software development and database construction 93

3.3 RESULTS AND DISCUSSION 94
  3.3.1 Distribution of *Fok*I restriction sites across the *E. coli* genome 94
  3.3.2 Amplification of indexed restriction fragments from *Fok*I-digested *E. coli* chromosomal DNA using pairs of phosphorylated indexers 95
  3.3.3 Coverage of the *E. coli* genome with *Fok*I restriction sites generating the cohesive end sequences CGCG and GCGC 98
  3.3.4 Evaluation of ligation and amplification reaction conditions for indexing-based profiling of bacterial genomes 101
    3.3.4.1 Rationale for the use of pooled NoP indexer mixes in ligation for bacterial fingerprinting 101
    3.3.4.2 Evaluation of ligation conditions for bacterial profiling 1: indexer concentration, DNA concentration, and ligase

concentration; comparison of T4 DNA ligase and Taq *DNA ligase* *103*

3.3.4.3 *Evaluation of ligation conditions for bacterial profiling 2: ligation temperature, ligation time and ligase concentration; comparison of T4 DNA ligase and* E. coli *DNA ligase* *109*

3.3.5 Amplification of IGP ligation reactions 113

3.3.5.1 *Background amplification patterns from ligations predicted to be devoid of legitimate indexing targets* *113*

3.3.5.2 *Comparison of amplification characteristics of* Taq *DNA polymerase and* PfuTurbo™ *DNA polymerase for IGP applications* *116*

3.3.6 Preliminary evaluation of correlation between predicted and detected indexing target fragments for indexed genomic profiling approaches... 117

3.3.6.1 *Software development and database construction* *117*

3.3.6.2 *Use of EcoliDB v1.0 to identify predicted indexable* FokI *restriction fragments in* E. coli *chromosomal DNA* *118*

3.3.6.3 *Preliminary experimental evaluation of correlation between predicted and detected indexing target fragments* *120*

3.3.7 Indexed genomic profiling of *E. coli* K12 strain MG1655 123

3.3.8 Differentiation of *E. coli* strains by IGP 125

3.3.8.1 *Comparison of four* E. coli *strains using a small subset of IGP reactions* *125*

3.3.8.2 *Indexed genomic profiling of three* E. coli *laboratory strains* *126*

3.3.8.3 *Indexed genomic profiling of* Staphylococcus *reference species and clinical isolates* *130*

3.4 DISCUSSION 130

3.4.1 Summary of indexed genomic profiling studies 130

3.4.2 Indexed genomic profiling as a potentially definitive method for bacterial genotyping 132

3.4.3 **Potential applications of indexed genomic profiling** 134

3.4.3.1 Clinical and epidemiological studies of pathogenic Staphylococcus *species and strains* *134*

3.4.3.2 Studies of microbial community diversity *136*

4 Chapter IV: Global Gene Expression Profiling of *Saccharomyces cerevisiae* by 3'-end cDNA Indexing 139

4.1 INTRODUCTION 139

Differential display *139*

Serial analysis of gene expression *140*

cDNA microarrays *140*

3'-end cDNA indexing *141*

4.1.1 ***Saccharomyces cerevisiae*** **as a model organism for transcriptomics** 142

4.1.2 Global gene expression profiling of *S. cerevisiae* by 3'-end cDNA indexing 143

4.1.3 ABSTRACT: 147

4.2 MATERIALS AND METHODS 149

4.2.1 ***S. cerevisiae*** **strains and growth conditions** 149

4.2.1.1 Environmental Condition I: Glucose as a Carbon Source *149*

4.2.1.2 Environmental Condition II: Galactose as a Carbon Source *149*

4.2.1.3 Environmental Condition III: Response of Mating Type a Yeast to Mating Factor α *149*

4.2.1.4 Environmental Condition IV: Response of Yeast to High Osmolarity *149*

4.2.2 RNA preparation from *S. cerevisiae* by phenol/freeze method 150

4.2.3 Isolation of mRNA populations from yeast total RNA 151

4.2.4 cDNA synthesis: first-iteration protocol 153

4.2.4.1 First strand cDNA synthesis by RETROscript method using anchored poly(T)$_{16}$-V primers ... 153

4.2.4.2 Second strand cDNA synthesis by modified Klenow method ... 154

4.2.5 Digestion of double-stranded cDNA populations by *Fok*I restriction endonuclease ... 155

4.2.6 Amplification of biotinylated indexed cDNA fragments from *Fok*I-digested cDNA populations with BamCC and TF primers following Dynabead extraction ... 155

4.2.7 Modified protocol for ds-cDNA synthesis from *S. cerevisiae* mRNA populations ... 156

4.2.7.1 cDNA synthesis using GCRichPoly(T)$_{16}$-V primers and SuperScript™ System ... 156

First-strand cDNA synthesis using GCRichPolyT$_{16}$-V primers ... 157

Second-strand cDNA synthesis ... 157

Second-strand synthesis controls ... 158

4.2.7.2 Fok*I digestion of ds-cDNA populations primed with GCRichPoly(T)$_{16}$-V: ... 158*

4.2.8 Optimized ligation and amplification conditions for 3'-end cDNA indexing 158

4.2.9 3'-end cDNA indexing data acquisition and analysis by automated DNA sequencing instrumentation ... 160

4.3 RESULTS ... 161

4.3.1 Selection of Test Fragment (TF) target sequences and design of TF primer pairs ... 161

4.3.2 Use of Test Fragments as controls ... 162

4.3.2.1 Amplification of Test Fragments as controls for second-strand cDNA synthesis quality ... 162

4.3.2.2 Amplification of Test Fragments as control for ds-cDNA Fok*I digest quality ... 166*

4.3.3 Use of *Fok*I-digested Test Fragments as artificial target fragments for DNA indexing 166

4.3.3.1 Preparative amplification and Fok*I digestion of Test Fragment amplicons 167*

4.3.3.2 Initial indexing and amplification of Fok*I-digested TF amplicons 167*

4.3.3.3 Evaluation of minimum template concentrations using Fok*I-digested TF amplicons as models for 3'-end cDNA fragments ... 168*

4.3.4 Amplification of indexed *Fok*I-digested TF restriction fragments from *Fok*I-digested yeast cDNA populations using BamCC and TF primers. 173

4.3.4.1 Amplification of indexed Fok*I-digested TF restriction fragments directly from* Fok*I-digested yeast cDNA populations 173*

4.3.4.2 Amplification of biotinylated indexed cDNA fragments from Fok*I-digested cDNA populations with BamCC and TF primers following purification by streptavidin-coated paramagnetic beads. 174*

4.3.5 Evaluation of Dynabead wash regimens 178

4.3.5.1 Evaluation of wash protocols to prevent non-specific binding of nonbiotinylated DNA fragments to streptavidin-coated paramagnetic beads 178

4.3.5.2 Ligation and PCR of biotinylated pUC19 fragments for use in Dynabead wash regimen evaluations 179

4.3.5.3 Identification of wash regimens eliminating nonspecific DNA binding to streptavidin-coated paramagnetic beads 181

4.3.6 Attempted amplification of biotinylated indexed cDNA fragments from *Fok*I-digested cDNA populations with BamCC and anchored poly(T) primers following Dynabead extraction 183

4.3.7 Design, construction and use of FakePolyT model fragments in optimization of poly(T) priming strategies 187

4.3.7.1 Designing the BigPolyT Indexer for construction of FakePolyT fragments ... 187

4.3.7.2 Approach 1: Ligation of BigPolyT Indexer to pUC19 A & B fragments ... 187

4.3.7.3 Comparison of amplification efficiency of 3'-primer sets Poly(T)$_{35}$-V and Poly(T)$_{16}$-V ... 189

4.3.7.4 Design of GCRichPoly(T)$_{16}$-V primer ... 190

4.3.7.5 Approach 2: Construction of a DNA target amplicon primed at both ends by poly(T) primers ... 190

Amplification of AB^{++} construct with anchored poly(T) primers 193

Digestion of poly(T)-primed AB^{++} amplicons ... 193

Ligation of BamCC NoP indexers to Fok*I-digested poly(T)-primed AB^{++} constructs for completion of FakePolyT assembly ... 195*

4.3.8 Use of FakePolyT fragments in development of cDNA indexing protocols ... 196

4.3.8.1 Comparison of efficiency of Poly(T)$_{35}$-V and GCRichPoly(T)$_{16}$-V primers for indexed 3'-end cDNA fragment amplification ... 196

4.3.8.2 Efficiency of 3'-end cDNA fragment amplification from Dynabead-bound biotinylated templates ... 198

4.3.8.3 Evaluation of PCR cycling parameters for 3'-end cDNA indexing ... 198

4.3.8.4 Estimation of indexed 3'-end cDNA template requirement for amplification with GCRichPoly(T)$_{16}$-V and BamCC ... 200

4.3.9 GC-rich anchored poly(T)-primed synthesis and quality assurance of ds-cDNA populations from total cellular mRNA isolated from *S. cerevisiae* cultures grown under different environmental conditions. 202

4.3.10 Software development and database construction ... 203

4.3.11 Differential expression of *GAL1* and of *BOP3* reported by 3'-end cDNA indexing 205
4.3.12 Conditions providing high ligation fidelity by *Taq* DNA ligase 210
4.3.13 Reporting by 3'-end cDNA indexing of differential gene expression in yeast in response to saline stress 210
4.3.14 3'-end cDNA indexing data acquisition and analysis by automated DNA sequencing instrumentation 215
4.3.15 Evaluation of global gene expression profiling in *S. cerevisiae* by 3'-end cDNA indexing 217
4.3.15.1 Reproducibility of 3'-end cDNA indexing profiles of gene expression in S. cerevisiae *218*
4.3.15.2 Comparison of cDNA indexing gene expression profiles generated from saline-treated and untreated yeast cultures 221
4.3.15.3 Evaluation of 3'-end cDNA indexing of S. cerevisiae *for global gene expression profiling: progress summary 224*

4.4 DISCUSSION 226
4.4.1 Development of 3'-end cDNA indexing protocols for global gene expression profiling in *S. cerevisiae*: summary 226
4.4.2 Future Research: addressing the challenges to 3'-end cDNA indexing identified in this investigation 228
4.4.3 Future Research: exploitation of the "C_0t effect" in future cDNA indexing applications 230

5 Chapter V - Conclusions and Future Work 232

5.1 CHAPTER SUMMARIES AND CONCLUSIONS 232
5.1.1 Development of DNA Indexing Strategies for Directed Mapping and Sequencing 232

5.1.2 Bacterial Strain and Species Differentiation by Indexed Genomic Profiling (IGP)....234

5.1.3 Global Gene Expression Profiling of *Saccharomyces cerevisiae* by 3'-end cDNA Indexing....235

5.2 Future Developments: Indexing-Directed Bacterial Genomics (IDBG)......236

5.2.1 Identification of Type IIS cohesive end sequences by multiplex indexing236

5.2.2 Indexing sequence-tagged site (iSTS) mapping....240

5.2.3 Indexing-Directed Bacterial Genomics (IDBG): an efficient non-cloning method for the directed mapping and sequencing of prokaryotic genomes....242

5.2.3.1 Sfi-series indexer sets for genomic mapping and sequencing.....242

5.2.3.2 Classification and isolation of Sfi*I fragments based on cohesive end sequence....246*

5.2.3.3 Assembly of the iSTS physical map....247

5.2.3.4 Fok*I-based subcontig mapping and sequencing of* Sfi*I fragments....251*

5.2.3.5 Sfa*NI-based subcontig mapping and sequencing of* Sfi*I fragments....253*

5.2.3.6 Assembly of sequenced SfiI fragment contigs along the iSTS physical map for complete genome sequence construction........254

6 Bibliography....255

List of Figures

Figure 2.1 Schematic flow diagram for DNA indexing..............................24

Figure 2.2 *Fok*I restriction map of plasmid pUC19................................26

Figure 2.3 Features of the α phosphorylated indexer set and the γ nonphosphorylated indexer set..45

Figure 2.4 Digestion of pUC19 by *Fok*I restriction endonuclease.................49

Figure 2.5 Target-specific ligation of indexers permits amplification of the indexed fragment..52

Figure 2.6 Single-primer P/P indexing of pUC19 by α indexers..................55

Figure 2.7 P/NoP indexing circumvents the collateral amplification of repeated-end fragment classes..59

Figure 2.8 Determination of optimal ligation conditions for DNA indexing....62

Figure 2.9 Production and structure of indexing PD artifact from double-primer P/NoP indexing of pUC19..67

Figure 2.10 Features of the Bam phosphorylated indexer set and the BamCC nonphosphorylated compound indexer set..............................70

Figure 2.11 Compound-primer P/NoP strategy...73

Figure 2.12 Compound-primer P/NoP indexing of pUC19 fragments............76

Figure 2.13 Indexing of pUC19 *Fok*I fragments for sequencing template production...78

Figure 2.14 Amplification of B_a and B_b sequencing templates for a *Fok*I/*Sfa*NI double restriction digest..82

Figure 2.15 Index map of pUC19 sequencing template coverage..................86

Figure 2.16 Complete indexing-based sequencing contig of pUC19...............87

Figure 3.1 Amplification of indexed restriction fragments from *Fok*I-digested *E. coli* chromosomal DNA 96

Figure 3.2 Coverage of the *E. coli* genome with *Fok*I restriction sites generating the cohesive end sequences CGCG and GCGC.........100

Figure 3.3 Evaluation of ligation conditions for bacterial profiling 1: indexer concentration, DNA concentration and ligase concentration; comparison of T4 DNA ligase and *Taq* DNA ligase....................106

Figure 3.4 Evaluation of ligation conditions for bacterial profiling 2: ligation temperature, ligation time and ligase concentration; comparison of T4 DNA ligase and *E. coli* DNA ligase...................................110

Figure 3.5 Background amplification patterns from ligations predicted to be devoid of legitimate indexing targets....................................115

Figure 3.6 Correlation between predicted and detected indexing target fragments..122

Figure 3.7 Indexed genomic profiling of *E. coli* K12 MG1655....................124

Figure 3.8 Comparison of four *E. coli* strains using a small subunit of IGP reactions..127

Figure 3.9 Indexed genomic profiling of three *E. coli* laboratory strains......128

Figure 3.10 Indexed genomic profiling of *Staphylococcus* reference species and clinical isolates...131

Figure 4.1 Expression profiling of *S. cerevisiae* by 3'-end cDNA indexing....144

Figure 4.2 Design of Test Fragment target sequences.............................164

Figure 4.3 Use of Test Fragments as controls for second-strand cDNA synthesis and for ds-cDNA *Fok*I digest quality........................165

Figure 4.4 Initial indexing and amplification of *Fok*I-digested TF amplicons...169

Figure 4.5 Estimation of minimum tempate concentration requirements using *Fok*I-digested TF amplicons...171

Figure 4.6 Amplification of biotinylated indexed pUC19 fragments from *Fok*I-digested yeast cDNA populations.......................................175

Figure 4.7 Amplification of biotinylated indexed pUC19 fragments for use in evaluating Dynabead wash regimens....................................180

Figure 4.8 Evaluations of wash regimens for reduction of nonspecific DNA binding to streptavidin-coated paramagnetic beads.................184

Figure 4.9 **Design, construction and use of FakePolyT model fragments in optimization of poly(T) priming strategies............................188**

Figure 4.10 **Construction of FakePolyT amplicons primed at both ends by poly(T) primers..192**

Figure 4.11 **Indexing of *Fok*I-digested poly(T)-primed AB^{++} amplicons 194**

Figure 4.12 **Comparison of efficiency of Poly(T)$_{35}$-V and GCRichPoly(T)$_{16}$-V primers for amplification of indexed FakePolyT fragments........197**

Figure 4.13 **Amplification of biotinylated indexed artificial 3'-end cDNA fragments following streptavidin capture..............................199**

Figure 4.14 **Coarse estimation of amplification efficiency for indexed GC16-tailed 3'-end cDNA fragments..201**

Figure 4.15 **Evaluation of GC-rich anchored poly(T)-primed cDNA synthesis quality and *Fok*I digest quality using TF primer pairs..............204**

Figure 4.16 **Differential expression of *GAL1* and *BOP3* reported by 3'-end cDNA indexing...208**

Figure 4.17 **Differential expression in yeast in response to saline stress.........212**

Figure 4.18 **Reproducibility of 3'-end cDNA indexing profiles of gene expression in *S. cerevisiae* 1: parallel indexing iterations of an individual *Fok*I-digested cDNA population...219**

Figure 4.19 **Reproducibility of 3'-end cDNA indexing profiles of gene expression in *S. cerevisiae* 2: parallel indexing iterations of parallel *Fok*I-digested cDNA populations..220**

Figure 5.1 **Schematic flow diagram of Indexing-Directed Bacterial Genomics...243**

Figure 5.2 **Features of the three proposed Sfi indexer sets.......................244**

List of Tables

Table 1.1 Relation of restriction fragment cohesive end length to indexing system complexity……3
Table 1.2 Comparison of indexing system complexity for candidate restriction endonucleases……3
Table 1.3 Estimated effect of restriction endonuclease selection on indexing system complexity for two genome sizes……7
Table 2.1 pUC19 target fragment end sequences and fragment sizes……27
Table 2.2 Oligo sequences employed in DNA indexing investigations……32
Table 3.1 Anticipated targeting of ***E. coli*** chromosomal ***Fok*I** fragments by selected indexer combinations……121
Table 4.1 Sequences of Test Fragment primer pairs……163
Table 4.2 Description of Test Fragments……163
Table 4.3 Identification of differentially-expressed transcripts by 3'-end cDNA indexing……209
Table 4.4 Identification of indexed cDNA transcripts expressed in response to saline stress……213
Table 4.5 Reporting of differential gene expression in response to saline stress by 3'-end cDNA indexing……223
Table 5.1 Multiplex mix compositions for cohesive end identification……238
Table 5.2 Decoding of binary multiplex signals to identify indexed cohesive end sequences……238

List of Abbreviations

%GC	percent guanine/adenine content
64-mix	one of 64 indexer pools, each containing four NoP indexers
AE	sodium acetate - EDTA
AFLP	amplified-fragment length polymorphism
AMP	adenosine monophosphate
Amp-LB	ampicillin-containing Luria broth
AP-PCR	arbitrarily-primed polymerase chain reaction
ARDRA	amplified ribosomal DNA restriction analysis
ASCII	American Standard Code for Information Interchange
ATAC-PCR	adaptor-tagged competitive PCR
ATCC	American Type Culture Collection
ATP	adenosine triphosphate
BAC	bacterial artificial chromosome
bp	basepair
BSA	bovine serum albumin
cDNA	complementary DNA
CIAA	chloroform-isoamyl alcohol
D-HBV	duck hepatitis B virus
ddH_20	deionized distilled water
DEPC	diethyl pyrocarbonate
DNA	deoxyribonucleic acid
dNTPs	deoxynucleotide triphosphates
DTT	dithiothreitol
EDTA	disodium ethylenediamine tetraacetic acid
EtBr	ethidium bromide
EtOH	ethanol
FAM	5-carboxyfluorescein
g	gravitational force

HMG-14	high-mobility-group nonhistone structural protein 14
IDBG	indexing-directed bacterial genomics
IGP	Indexing Genomic Profiling
IP	interrupted palindrome
IPTG	isopropylthio-β-D-galactoside
iSTS	indexing sequence-tagged site
JOE	6-carboxy-4',5'-dichloro-2',7'-dimethoxyfluorescein
kb	kilobasepairs
LB	Luria broth
LCR	ligase chain reaction
Mb	megabasepair
MOPS	(morpholino)propanesulfonic acid
MPC	magnetic particle concentrator
mRNA	messenger RNA
MRSA	methicillin-resistant *Staphylococcus aureus*
MW	molecular weight
NaOAc	sodium acetate
NAD^+	nicotinamide adenine dinucleotide
NbFeB	neodymium iron boron
NH_4OAc	ammonium acetate
NoP indexer	nonphosphorylated (5'-hydroxylated) indexer
nt	nucleotide
oligo	oligonucleotide
ORF	open reading frame
P-indexer	5'-phosphorylated indexer
P/C/IAA	phenol-chloroform-isoamylalcohol
PCR	polymerase chain reaction
PTC	programmable thermal controller
PD	primer-dimer
PEG	polyethylene glycol

PFGE	pulsed-field gel electrophoresis
poly(A)	polyadenylated
PPi	inorganic pyrophosphate
PRE	pheromone-responsive element
RAPD	random amplification of polymorphic DNA
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
ROX	6-carboxy-X-rhodamine
rpm	revolutions per minute
rRNA	ribosomal RNA
SAGE	serial analysis of gene expression
SDS	sodium dodecylsulfate
SSC	standard saline citrate
T-RFLP	terminal restriction fragment length polymorphism
TAE	Tris-acetate EDTA
TAMRA	6-carboxytetramethylrhodamine
TE	Tris-EDTA
TF	Test Fragment
Tris	tris(hydroxymethyl)aminomethane
Tris-HCl	Tris-hydrochloric acid
Type IIS	Type II-"shift"
U	enzymatic functional units
U_L	cohesive-end ligation units
UV	ultraviolet
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactoside
YAC	yeast artificial chromosome
YPD	yeast extract-peptone-dextrose
YPG	yeast extract-peptone-galactose

1 Chapter I: Introduction to DNA Indexing

1.1 DNA indexing

1.1.1 Historical context

The foundation for much future research, the determination of the three billion basepairs of human genomic DNA sequence has presented a formidable analytical challenge. Mapping the genome using clonal libraries, sequencing genomic fragments from these libraries, and analyzing the resulting data were daunting tasks using the approaches available at the outset of this international undertaking [1-4]. Contemporary techniques such as yeast artificial chromosome (YAC) library preparation and primer walking were laborious, required high redundancy, and were subject to the problems intrinsic to cloning including selective fragment loss, chimera formation, mutation and sequence rearrangement [1, 5-8]. While these standard methods were being employed in efforts to elucidate the physical structure of the human genome, and the genomes of numerous model organisms, the search for greater efficiency and applicability spurred research in the development of novel technologies [9-15]. It was clear that further study of these genomes would be facilitated by an ordered approach to complete sequence analysis and directed-search capabilities for genes and controlling elements [16-21]. Also, in anticipation of an increased emphasis on comprehensive investigations of functional genomics, new approaches to functional analysis, specifically the measurement of expression profiles of genes, needed to be developed. To match the scope and power of structural genome investigations, methods to record the expression states of large numbers of expressed genes needed to be efficient, high throughput, and be amenable to automation. In this context, the DNA indexing approach developed by Unrau and Deugau [22] was intended to fulfill the identified requirements by providing a core set of concepts and techniques forming the basis for a suite of powerful molecular biology applications.

1.1.2 Principles of DNA indexing

DNA indexing provides an ordered, non-cloning approach for defining, amplifying and isolating specific DNA fragments from complex genomic digests. Type IIS (or Type II-"shift") and Interrupted Palindrome (IP) restriction endonucleases cleave DNA at defined distances from their recognition sites, generating fragments with non-identical cohesive ends that are not predictable from the recognition sequence but are specific for any particular cut site [23, 24]. For example, the Type IIS restriction enzyme *Fok*I recognizes the sequence GGATG and cuts 9 bases downstream on the coding strand and 13 bases downstream on the complementary strand, leaving a 4-base 5'-cohesive end [25]. The cohesive ends generated by Type IIS and IP endonucleases thus bear information in their single-stranded overhanging sequences. These informative cohesive ends can be modified by ligation of specific indexing oligonucleotides (or *indexers*), to which various functional groups can be added to allow cohesive-end identification and fragment manipulation. Indexers consist of two annealed synthetic oligonucleotides: an indexing strand bearing a common primer-binding sequence and a variable 4-nucleotide 5'-cohesive end; and a complementary common primer strand. Selection of the appropriate indexer for, and its ligation to, each cohesive end sequence of a targeted restriction fragment incorporates a common-primer-binding sequence onto each end of that fragment, allowing specific amplification of the indexed target fragment in isolation from other non-indexed restriction fragments present in the same reaction.

DNA fragments from a Type IIS or IP restriction digest may be classified or defined by their cohesive end sequences (TABLE 1.1). For cohesive end sequences N bases in length, the number of possible sequences (M) is 4^N. Each fragment in such a digest has two independently indexable cohesive ends, each end having M possible sequences. A fragment is classified by the particular sequences of the two indexers required to base pair and ligate to the fragment's cohesive ends in order to permit target amplification. For illustrative purposes, we can consider the case of a restriction enzyme which produces single-base cohesive ends. Digestion of DNA with the Type

TABLE 1.1: Relation of restriction fragment cohesive end length to indexing system complexity.

Example Enzyme	Cohesive End Length	Cohesive End Classes	Fragment Classes
*Scr*FI	1	4	10
*Drd*I	2	16	136
*Sfi*I	3	64	2 080
*Fok*I	4	256	32 896
*Tsp*RI	4.5*	512	65 536
*Hga*I	5	1 024	524 800

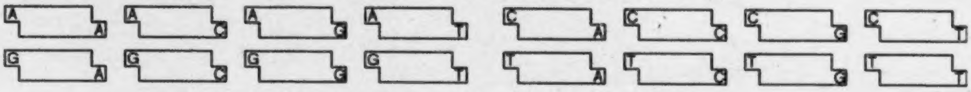
* This enzyme effectively has a 4.5 bp recognition sequence and 4.5 informative nucleotides on its 9-ntcohesive end sequence.

TABLE 1.2: Comparison of indexing system complexity for candidate restriction endonucleases.

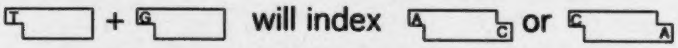
Enzyme	Recognition Sequence	Cohesive End Length (nt)	Recognition Sequence Length (bp)	Fragment Length (bp)	Fragment Classes
5'-protruding cohesive ends:					
*Bbs*I	GAAGACN(2/6)	4	6	2 048	32 896
*Bbv*I	GCAGCN(8/12)	4	5	512	32 896
*Bsa*I	GGTCTCN(1/5)	4	6	2 048	32 896
*Bsm*AI	GTCTCN(1/5)	4	5	512	32 896
*Bsm*BI	CGTCTC(1/5)	4	6	2 048	32 896
*Bsm*FI	GGGAC(10/14)	4	5	512	32 896
*Bsp*MI	ACCTGCN(4/8)	4	6	2 048	32 896
*Ear*I	CTCTTCN(1/4)	3	6	2 048	2 080
*Eco*31I	GGTCTCN(1/4)	3	6	2 048	2 080
*Esp*3I	CGTCTCN(1/5)	4	6	2 048	32 896
*Fok*I	GGATGN(9/13)	4	5	512	32 896
*Hga*I	GACGCN(5/10)	5	5	512	524 800
*Ksp*632I	CTCTCN(1/4)	3	5	512	2 080
*Sap*I	GCTCTTCN(1/4)	3	7	8 192	2 080
*Sfa*NI	GCATCN(5/9)	4	5	512	32 896
*Sts*I	GGATGN(10/14)	4	5	512	32 896
3'-protruding cohesive ends:					
*Alw*NI	CAGNNN/ CTG	3	6	4 096	2 080
*Apa*B1	GCANNNNN/ TGC	5	6	4 096	524 800
*Bgl*I5	GCCNNNN/ NGGC	3	6	4 096	2 080
*Bpu*10I	CCTNACG(5/2)	3	6	2 048	2 080
*Bsl*I	CCNNNNN/ NNGG	3	4	256	2 080
*Bst*API	GCANNNN/ NTGC	3	6	4 096	2 080
*Bst*XI	CCANNNNN/ NTGG	4	6	4 096	32 896
*Dra*III	GAGNNN/ GTG	3	6	4 096	2 080
*Mwo*I	GCNNNNN/ NNGC	3	4	256	2 080
*Pfl*MI	CCANNNN/ NTGG	3	6	4 096	2 080
*Rle*AI	CCCACAN(12/9)	3	6	2 048	2 080
*Sfi*I	GGCCNNNN/ NGGCC	3	8	65 536	2 080
*Tsp*RI	NNCA$^{\mathrm{C}}_{\mathrm{G}}$TGNN/	4.5*	4.5*	1 024	65 792

* This enzyme generates cohesive ends 9 nt in length, with an information content equivalent to 4.5 nt.

IP restriction endonuclease *Scr*FI (recognition sequence 5'-CC^{V}NGG-3') will generate the following types of single-base cohesive-ended fragments:

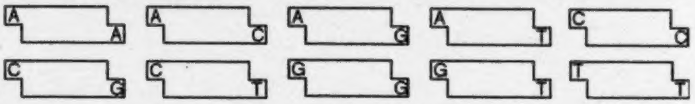


It is not possible to determine *a priori* a fragment's orientation on the basis of the indexers required to amplify that fragment: that is, indexing does not distinguish between a fragment amplified with indexer T on the left and indexer G on the right, and a fragment amplified with indexer G on the left and indexer T on the right:



As the subsets generated by indexing do not differentiate between fragment orientations for fragments amplified with the same two indexers, there are M x [(M + 1)/2] identifiable classes of fragments as defined by end sequence alone.

The number of fragment classes into which a Type-IIS restriction digest of complex DNA may be resolved is a fundamental parameter of any indexing system. The more discrete indexing fragment classes generated by a particular Type IIS restriction enzyme, the further subdivided any particular restriction fragment population will be, and the fewer indexable restriction fragments will fall into any one class, thus providing greater resolving power for indexing systems based on larger fragment class numbers. In the case of fragment mixtures with single-base cohesive ends, such as digests generated by *Scr*FI, there are ten fragment classes:



Indexing systems based on cohesive ends of two nucleotides permit resolution of digest fragments into a mere 136 classes. In contrast, enzymes that generate cohesive ends bearing three informative nucleotides permit the characterization of 2080 indexing fragment classes on the basis of end sequence alone. While this is clearly an improvement, it is still a small number of classes when the total number of restriction fragments generated (and thus the average number of fragments per class) is considered for genomes of moderate complexity. For indexing systems based on enzymes which generate 4-nt informative cohesive ends, there are 256 cohesive-end classes (one for each possible tetranucleotide sequence) and therefore 32 896 fragment classes based on end sequence discrimination alone. This large number of discrete fragment classes provides a more appropriate level of resolution for the analysis of complex genomes by DNA indexing.

For example, a restriction enzyme with a palindromic 6-bp recognition sequence will generate fragments of 4096 bp on average, if the %GC of the recognition sequence and the DNA substrate are assumed to be identical, and there is no relative hexanucleotide bias in the sequence of the DNA substrate. If the DNA substrate is a 5 Mb prokaryotic genome, complete digestion with such an enzyme would generate roughly 1200-1300 fragments. However, the fragments thus generated would be typically too long to allow efficient direct cycle sequencing of complete fragments at one time, and in some cases may be difficult to amplify without resorting to special long-fragment PCR conditions. If a restriction enzyme with a 4-bp recognition sequence were used (again assuming identical nucleotide bias between recognition sequence and DNA substrate), approximately 20 000 fragments averaging 256 bp in length would be generated. Digestion of the same genome by a restriction enzyme with a 5-bp recognition sequence would be expected to produce roughly 10 000 fragments with an average length of 512 bp. The size ranges of the fragments produced by digestion with each of these hypothetical enzymes are such that direct cycle sequencing of the fragments would be feasible in most cases.

The presence of 3-nt informative cohesive ends on digested fragments would permit digests generated by either the enzyme with a 4-bp recognition sequence or the enzyme with a 5-bp recognition site referred to in the previous examples to be efficiently characterized by indexing approaches (TABLE 1.2). In indexing analyses of the hypothetical genome mentioned above, use of an 4-bp-recognizing enzyme generating three informative nucleotides would produce roughly 20 000 fragments averaging 256 bp in length, falling into 2080 fragment classes where each end-sequence-defined fragment class would contain 10 fragments, on average. Alternately, use of an 5-bp-recognizing enzyme generating three informative nucleotides would produce roughly 10 000 fragments averaging 512 bp in length, falling into 2080 fragment classes where each end-sequence-defined fragment class would contain 5 fragments, on average.

The informative power of cohesive ends exploited by DNA indexing approaches can be seen by comparing the 3-nt system described above with the case of an indexing system based on enzymes generating 4-nt quasi-random cohesive ends (TABLE 1.3). In this instance the hypothetical 5 Mb bacterial genome is digested by a Type IIS restriction endonuclease that recognizes the same 5-bp sequence as the previous example but instead produces fragments with 4-nt informative ends. As before, such a digest would be expected to generate roughly 10 000 fragments averaging 512 bp in length. However, the fragments with 4-nt overhangs will fall into 32 896 end-sequence-defined fragment classes instead of only 2080 classes in the case of 3-nt overhangs. As a result, less than one-third of the fragment classes would be expected to be represented in the digest. Of the fragment classes which are represented, most will be represented by a single fragment. Of those fragment classes with more than one representative member in the digest, each representative member within that class is likely to be easily distinguishable from every other fragment in that class on the basis of size. This example demonstrates that indexing systems based on enzymes generating fragments bearing 4-nt informative cohesive ends provide the large number of discrete fragment classes suitable for the analysis of complex genomes, while requiring the synthesis of only a moderate number of oligonucleotides for complete description of those genomes by DNA indexing.

As indexers can be attached specifically to complementary sequences only, they can be used to classify DNA fragments according to cohesive end sequences, digestion enzyme and fragment length. This provides a non-cloning method to manipulate these fragments and separate them into well-characterized subsets. Following restriction endonuclease digestion of the DNA of interest, the specificity of the T4 DNA ligase reaction is used to selectively import appropriately selected indexers onto the cohesive ends of DNA fragments of a particular target fragment class. The primer-binding sequences borne by these two indexers permit specific amplification of the target class

TABLE 1.3: Estimated effect of restriction endonuclease selection on indexing system complexity for two genome sizes.

Genome	Indexing Endonuclease (e.g.)	Recognition Sequence (bp)	Cohesive End Length (nt)	Fragment Length (bp)	Fragment Classes	Fragments in Indexed Genome	Fragments per Class in Indexed Genome
Mammalian (3 Gb)	*Bsl* I	4	3	256	2 080	12 000 000	5 800
	None isolated	4	4	256	32 896	12 000 000	360
	None isolated	5	3	512	2 080	6 000 000	2 900
	Fok I	5	4	512	32 896	6 000 000	180
	Dra III	6	3	4 096	2 080	750 000	360
	Sts I	6	4	4 096	32 896	750 000	25
Bacterial (5 Mb)	*Bsl* I	4	3	256	2 080	20 000	10
	None isolated	4	4	256	32 896	20 000	unique; over 13 000 classes unrepresented
	None isolated	5	3	512	2 080	10 000	5
	Fok I	5	4	512	32 896	10 000	unique; over 22 000 classes unrepresented
	Dra III	6	3	4 096	2 080	1 250	unique; over 1 000 classes unrepresented
	Sts I	6	4	4 096	32 896	1 250	unique; over 31 000 classes unrepresented

(by PCR) from a population of fragments. Further manipulation of the target class, such as cycle sequencing or purification via biotin-streptavidin binding (using biotin-modified indexers), may then be performed as required by the indexing application of interest.

1.1.3 Estimation of system complexity for DNA indexing studies

It should be noted that several of the key parameters in the examples above are not actual values found in any particular indexing system, but instead represent approximations that rely for their validity on several underlying assumptions. Specifically, the values which rely on these assumptions are: the average length of fragment generated by a restriction enzyme with a particular number of bases in its recognition sequence; the average number of such fragments in a digest of a DNA molecule of a particular size; and the average numbers of fragments per indexing class to be found in such a digest.

The examples assume that the %GC of the DNA to be digested is identical to that represented in the recognition sequence of the enzyme used to perform the digestion. Further, they assume that the recognition sequence is represented roughly the same number of times in the complete DNA sequence as any other sequence of the same number of nucleotides, and that that number is large. The examples also assume that the distribution of any one of these short sequences is even throughout the genome. These assumptions may be summarized by the statement “the DNA sequence of the genome of interest is random and homogeneous”. The greater the discrepancy between each of these assumptions and the real-world system represented by a particular restriction enzyme with a particular recognition sequence and a particular DNA target, the greater the deviation between the estimated average lengths or numbers of fragments and the actual lengths or number of fragments generated.

In addition, the examples assume homogeneity of the genome not only on the large scale (over Mb regions) but on the very small scale as well. This assumption affects the estimation of fragment number per indexing class in two ways. The representation of each possible sequence of informative cohesive end is assumed to be roughly equal in number and in distribution throughout the genome, and to be independent of local %GC variations affecting distribution of enzyme recognition sites. In other words, if the (hypothetical) enzyme recognition sequence is 5'-GGCGCCN$_1$/N$_5$-3', the examples assume that the prevalence and distribution of the informative cohesive end ATTA will be quite similar to the prevalence and distribution of the informative cohesive end CGGG. In actual fact, the high %GC content of the recognition sequence suggests that it will occur more frequently in GC-rich regions of the genome than in regions with %GC content more closely matched to the overall genome average. Consequently, the likelihood that four of the next five bases from the recognition site are As or Ts is lower than the probability that four of the next five bases are Cs or Gs. The false assumption that recognition sites are evenly distributed throughout the genome thus has direct bearing on the assumption that the distribution of a particular informative end sequence is even throughout the genome, and affects indirectly the assumption that a given informative end sequence will be found with similar frequency as any other informative end sequence. Finally, variance between the %GC content of the enzyme's recognition sequence and the overall %GC content of the DNA to be digested will directly affect the accuracy of recognition site frequency estimates expected in the genome, as well as indirectly influencing the other parameters mentioned above.

1.2 Enzymes employed in DNA indexing systems

1.2.1 Type IIS and Type IP restriction endonucleases

The Type IIS restriction endonucleases are an unusual subclass of restriction enzymes that recognize a specific DNA sequence and cleave DNA nonspecifically a short distance away from that sequence. While in many cases they share similar cofactor requirements to other Type II restriction enzymes, their recognition sequences are asymmetric, and their cleavage sites are located a defined distance, up to 20 base pairs, to one side of their recognition sequence [23]. They represent roughly 5% of the characterized restriction-modification systems of all classes (I, II, and Type III) [26-29].

Commercially-available Type II restriction endonucleases of all subclasses represent over 190 different recognition sequence specificities [30, 31]. Of these, 58 are of the Type IIS or IP subclasses. In other words, these enzymes have cleavage sites with sequences that are not predictable from their recognition specificities. Forty-six of these restriction enzymes cleave DNA such that the quasi-random cohesive ends generated, if any, are three bases or less in length. The information content of these short cohesive ends permits only small numbers of indexing fragment classes to be defined, generating many fragments per class and thus complicating analysis of indexing reactions by agarose gel electrophoresis.

The restriction enzyme *Hga*I (recognition sequence 5'-GACGCN_5/N_{10}-3') generates 5-nt 5'-cohesive ends. This enzyme can generate 1024 different cohesive end sequences, leading to an excessively high number of discrete indexing fragment classes (524 800) and necessitating the synthesis of over a thousand oligonucleotides to generate even a single complete set of indexers. While this enzyme is an attractive option for some envisioned applications of DNA indexing (such as the mapping, genotyping and sequencing of "novel" large eukaryotic genomes), it was deemed impractical as a choice for the development of a DNA indexing model system due to the unique nature of its cohesive ends and the high number of oligo syntheses required for its use.

Another interesting case, the restriction endonuclease *Tsp*RI (recognition sequence 5'-NNCAXTGNN-3', where X is either C or G) cleaves at the 3'-end of its recognition sequence, leaving a 9-base 3'-cohesive end of which 4.5 bases are informative. That is, there is a set of 32 896 fragment classes defined by the informative end NNCACTGNN and another set of 32 896 fragment classes defined by the informative end NNCAGTGNN, for a total of 65 792 distinguishable fragment classes. A complete set of 512 indexers with 9-base 3'-cohesive ends would be required for an indexing system based on this enzyme. The value of such an indexing system would be limited by the unique nature of the cohesive ends generated by this enzyme, making it impractical to index a particular DNA molecule with an inappropriate number or distribution of *Tsp*RI sites simply by selecting a different enzyme with which to digest the genome. Primarily for this reason, *Tsp*RI would have been an inappropriate choice of enzyme to form the basis for developing and troubleshooting a robust and easily-adaptable indexing model system. However, development of an indexing system exploiting the unique features of the cohesive ends generated by *Tsp*RI presents a tantalizing prospect for future investigation, particularly from the perspectives of cohesive end stability and ligation fidelity.

Ten commercially-available restriction enzymes generate 4-nt informative cohesive ends. All but one of these are Type IIS enzymes that generate 4-nt 5'-overhangs. (The sole exception is the IP enzyme *Bst*XI, which generates 4-nt 3'-cohesive ends.) An indexing system based on the cohesive end pattern of one of these nine enzymes (i.e. the synthesis of sets of 256 indexers, each indexer bearing one of the 4-nt tetranucleotide sequences on its 4-base 5'-overhang) has the advantage of being immediately applicable to each of the eight other enzymes. (In other words, if the use of one enzyme to digest and permit indexing analysis of a particular DNA molecule is found to be impractical due to an inappropriate number or distribution of that enzyme's recognition sites, the problem may be solved simply by selecting another one of the remaining eight enzymes as the indexing endonuclease.) For these reasons, a DNA indexing system based on 4-base 5'-cohesive ends would permit complete model system development and optimization of a model system with manageable numbers of oligo syntheses, and allow flexibility in restriction endonuclease selection for DNA substrate compatibility.

Four of the 4-base 5'-cutters (*Bbs*I, *Bsa*I, *Bsm*BI, and *Bsp*MI) have nonpalindromic 6-bp recognition sites. All things being equal, these enzymes might be expected to generate fragments averaging 2024 bp in length. As we have seen, fragment sizes in this range are difficult to sequence directly by cycle sequencing, and may in some cases be difficult to amplify at all, particularly under multiplex PCR conditions. The recognition sites for *Bbv*I and *Bsm*I, though both 5 bp in length (5'-GCAGCN$_8$/N$_{12}$-3' and 5'-GGGACN$_{10}$/N$_{14}$-3', respectively), are composed of sequences that are 80% GC. It is likely that this extreme %GC requirement skews both the number and the distribution of restriction fragments generated from most target DNAs, while biasing the representation of particular cohesive end sequence in favour of sequences with greater that 50% GC. Finally, three Type IIS restriction endonucleases (*Bsm*AI, *Fok*I, and *Sfa*NI) recognize 5-bp sequences of 60% GC and generate 4-nt 5' informative cohesive ends.

1.2.2 T4 DNA ligase

Polynucleotide ligases play essential roles in DNA repair, DNA recombination and DNA replication. The enzyme catalyzes the conversion of the pyrophosphate linkage of a nucleotide cofactor into a phosphodiester bond at single-strand breaks between adjacent 5'-phosphate and 3'-hydroxyl termini in double-stranded DNA [32]. The ligation reaction consists of three successive nucleotidyl transfer reactions. First, the nucleotide cofactor "activates" DNA ligase through the formation of a covalent protein-AMP intermediate, releasing free PPi. The nucleotide is then transferred from the enzyme to the phosphorylated 5'-end of a nicked strand of DNA to produce an inverted 5'-5' pyrophosphate bridge. Finally, DNA ligase catalyzes a transesterification reaction through nucleophilic attack of the 3'-hydroxyl group of the adjacent strand, joining the nick and releasing free AMP [33].

Two classes of DNA ligases have been characterized on the basis of the nucleotide cofactor required for ligase function. Eubacterial DNA ligases require NAD^+ as cofactor, becoming adenylated through cleavage of NAD^+ and release of nicotinamide mononucleotide [32, 34]. Viral, archaeal and eukaryotic DNA ligases utilize ATP as cofactor, and bear significant sequence homology to each other, RNA ligases, tRNA ligases and eukaryotic mRNA "capping enzymes" [35]. The ATP-dependent DNA ligases are capable of ATP-dependent self-adenylation and AMP-dependent DNA relaxation, as well as the capacity to catalyze the ligation of nicked, blunt-ended or sticky-ended fragments [36]. The archetype of this enzyme class is the DNA ligase encoded by the bacteriophage T4.

T4 DNA ligase has been a vital part of the molecular biologist's enzymatic toolkit since Khorana and co-workers used the enzyme to assemble base-paired oligonucleotide segments for the *de novo* construction of duplex DNA corresponding to the complete sequence of yeast alanine tRNA sequence [37]. *In vivo*, the enzyme typically functions in a "conservative" mode, mediating the joining of adjacent DNA strand termini correctly base paired to a complementary strand in a manner that preserves the integrity of the DNA sequence. T4 DNA ligase is also capable *in vitro* of catalyzing template-independent joining of DNA molecules with fully base-paired (blunt) ends at high concentrations of enzyme and blunt-ended DNA substrates ("radical mode" ligase activity). The diverse ligation events which this enzyme is capable of mediating make it suitable for a wide range of *in vitro* DNA manipulation and cloning techniques. However, as is the case for any other tool, the utility of T4 DNA ligase for a particular application depends on an evaluation of the characteristics of the enzyme relative to the requirements of the task. An exploration of T4 DNA ligase's anticipated mechanism of action in the catalysis of an indexing ligation event assists in such an evaluation.

In a ligation reaction containing indexers and *Fok*I-digested DNA fragments, a T4 DNA ligase molecule is adenylated by the ATP provided in the ligation buffer, forming an "activated" protein-AMP complex. The high-energy bond required to synthesize the phosphodiester bond is thus stored in the ligase prior to encountering a polynucleotide chain. As the activated ligase comes into contact with either an indexer or a restriction fragment, it forms a complex with the DNA substrate. DNA-binding studies have shown that stability of the enzyme-DNA complex is dependent on the adenylation state of the enzyme [36]. Below a critical ATP concentration, only deadenylated ligase is present, forming a highly stable complex on the DNA substrate. Adenylated enzyme in contrast forms only a very transient complex with DNA. Models suggest that the adenylated ligase "scans" the DNA through a series of transient complexes, searching for a phosphorylated 5'-end [36]. Following transfer of the AMP moiety to the phosphorylated 5'-end, the now-deadenylated ligase stalls on the DNA until a suitable 3'-hydroxyl group from another DNA fragment or indexer becomes available to complete the reaction.

In a ligation event involving a DNA restriction fragment and a double-stranded indexing oligonucleotide, successful presentation of the 3'-hydroxyl group of the incoming indexer is determined by the geometry of the duplex formed by the base pairing of the cohesive ends of the indexer and the target fragment [38]. Disruption of that geometry, for instance by mismatched bases on the incoming cohesive end sequences, may alter the angle between the 3'-hydroxyl, the 5'-phosphate it attacks and the AMP leaving group, and thus alter the rate of ligation [39, 40]. When a suitable 3'-hydroxyl end is presented by an indexer with a well-matched cohesive end sequence, the transesterification reaction is completed quickly, releasing AMP and free ligase. In the presence of ATP the enzyme is rapidly re-adenylated to start a new cycle. An interesting feature of this model is the description of each indexing ligation event (attachment of one indexer to one cohesive-ended target fragment) as the result of two successive second-order reactions ([free ligase] x [free DNA] => [stable complex] x [free indexer] => [free ligase] x [ligated DNA]) rather than a single third-order reaction ([free ligase] x [free DNA] x [free indexer] => [free enzyme] x [ligated DNA]).

The fidelity of the ligation reaction, that is, the degree to which a ligase requires that adjacent strands on either side of a nick be accurately base paired to the complementary "template" strand, is of critical importance to indexing. The requirement for a specific angle of attack by the 3'-hydroxyl group of the incoming strand on the adenylated 5'-phosphate group of the cohesive end of the targeted ligase-DNA complex contributes to the fidelity of DNA ligases [38]. Overall ligation fidelity is also affected by DNA ligase's ability to make appropriate contacts with the DNA duplex in the region of the joining reaction and by a reduction on the stability of the duplex by mismatches [38]. In general, DNA ligases favour the geometry provided by the correct base pairing of cohesive end sequences. However, it is well recognized that under forcing conditions *in vitro*, T4 DNA ligase can be induced to tolerate mismatches between the template strand and the nicked strands it joins together [41-43]. Certain mismatches, such as the G:T wobble base pair, cause only minimal disruption of duplex structure [44], and are tolerated by the DNA ligase reaction under certain conditions [39]. Harada and Orgel have demonstrated that under specific conditions efficient ligation may proceed despite mismatches at or near the junction at the 5'-terminus of a cohesive end. In some cases of specific artificial substrates (e.g. tight hairpin structures), complementary template and substrate sequences are not optimal for ligation [45]. Although it has also been demonstrated that completely complementary sequences are in general strongly favoured over mismatches in the ligation of more conventional substrates (see below), it is clear that ligation under non-stringent conditions is not proof of extensive complementarity between cohesive end sequences.

Historically, the utility of the DNA joining reaction mediated by T4 DNA ligase has been most often applied to the construction of recombinant DNA for cloning purposes. Most users interested in refinements of the ligation reaction seek to maximize the rate at which DNA vector and insert fragments are joined *in vitro* in order to obtain optimal cloning efficiency. The requirement that the cohesive ends present on the DNA fragments of interest be correctly base paired with one another is less relevant than achieving assembly of the desired recombinant DNA construct. As a result, much of the literature describing methods by which ligase performance might be improved for molecular biology applications employ ligation efficiency rather than ligation fidelity as the criterion for evaluating enhancement. These approaches include alterations to the incubation temperature, modifications to the concentrations of substrate or enzyme employed, or the addition of forcing agents which increase the apparent rate of ligation and thus facilitate basepairing-independent ligation events.

Studies of DNA ligase reaction rate relative to incubation temperature for a particular substrate set, which might be expected to reveal a bell-shaped curve centered on a particular temperature optimum as for other enzymes, are complicated by the fact that the duplex DNA substrate is also temperature-sensitive [38]. The optimal temperature for the action of T4 DNA ligase on physiological substrates is 37°, but in instances involving the joining of separate DNA fragments, the stability of the substrates' interaction at their termini becomes an important parameter in both the ligation rate and the extent of completion of the ligation reaction [46]. Melting temperatures of 4-nt cohesive ends nominally range between 8° and 16°C, depending on %GC content and nearest-neighbour interactions within the end sequence (Oligo Analyzer 2.0, Integrated DNA Technologies Inc., Coralville IO). As mismatched oligonucleotide duplexes are denatured at lower temperatures than accurately base paired duplexes of equal length, selection of an appropriate incubation temperature permits the discrimination of perfectly base-paired ends from mismatched end sequences and favours the ligation of the correctly matched complex over the mismatch. *In vitro*, the maximum ligation rate for short duplexes by T4 DNA ligase occurs at a slightly higher temperature than the measured thermal denaturation point of the duplex [47], suggesting that the ligase acts very rapidly on short-lived, base-paired interactions between cohesive-ended fragments, or that it stabilizes the interacting DNA substrates before joining them. The temperature at which 50% of the maximum reaction is still observable increases in relation to the %GC content of the cohesive end sequence [46]. In general, low temperatures reduce ligase activity, whereas elevated temperatures reduce cloning efficiencies by melting annealed DNA overhangs and increase overall molecular motion in the ligation reaction [48]. As a consequence, ligation of cohesive DNA ends for cloning purposes is normally performed at relatively low temperatures (12-16°C) to provide a balance between enzyme activity and stability of annealed DNA overhangs [49, 50].

The unusual ability of T4 DNA ligase to mediate the joining of two blunt-ended DNA fragments under extremely forcing conditions [51, 52], without requirement for template annealing or base pairing between incoming strands, has been exploited in a variety of widely-used DNA manipulation techniques [53-56]. Ligation of blunt-ended DNA fragments by T4 DNA ligase is highly inefficient compared to cohesive-end ligation, and is normally performed at lower temperatures (4–14°C) using much higher concentrations of both DNA substrates and of T4 DNA ligase (see Sobczak and Duguet, 1988 for review [57]). Efforts to enhance the performance of blunt-end ligation reactions involve the addition of agents which promote the rate of basepairing-independent ligation events. High concentrations of macromolecular crowding agents such as polyethylene glycol [52, 58, 59] or hexamine cobalt chloride [60] facilitate non-specific joining of DNA molecules through a volume exclusion mechanism, mimicking the effects of much higher DNA concentrations than are actually present in the reaction [61]. Positively-charged polyamines such as spermine, spermidine or protamine affect DNA-ligase interaction by neutralizing the phosphate groups of the DNA substrates [62, 63]. The addition of chromatin structural proteins such as histone H1 [57] and high-mobility-group nonhistone structural protein HMG-14 [63] may favour non-specific "radical" ligation events through the ability of these proteins to bring different DNA segments into close proximity, increasing the probability of the joining of these fragments by DNA ligase. The presence of forcing agents, and the use of high concentrations of enzyme and DNA substrate, perturb the fidelity requirements of DNA ligase and permit joining of diverse forms of DNA substrates. Under extreme conditions similar to those required for blunt-end ligation, T4 DNA ligase can even be induced to promote the ligation of blunt ends to ends bearing a 2- to 4-nt single-stranded protrusion [64, 65]. Although T4 DNA ligase is capable of such "radical mode" ligation events, the enzyme strongly favours "conservative mode" activities when non-forcing ligation conditions are selected, to the virtual exclusion of non-conservative events.

As previously stated, most applications of T4 DNA ligase do not require particularly stringent fidelity of ligation; rather, it is the efficiency of incorporation of a DNA fragment of interest into a linearized vector molecule, for example, that is the goal. In the context of these applications, the requirement that the cohesive ends present on the DNA fragments of interest be correctly base paired is less relevant than achieving assembly of the desired recombinant DNA construct. For the utility of T4 DNA ligase in the context of a DNA indexing system, however, the reverse is true. The specificity of fragment class amplification is directly related to the fidelity of the ligation reaction employed to import the indexing primers onto target fragments of that class alone. Clearly, successful application of T4 DNA ligase to the development of a DNA indexing system requires the establishment of a set of reaction conditions that favour high ligation fidelity at reasonable ligation rates, rather than emphasizing maximum ligation efficiency as the sole criterion for optimization.

A number of groups have presented evidence that under appropriate (non-forcing) reaction conditions the joining of oligonucleotides by T4 DNA ligase is highly sensitive to the presence of mismatched base pairs on either side of the joining site [66-68]. Elevated incubation temperatures (>30°C) greatly enhances the specificity of the ligation reaction [66]. The presence of a single mismatched base pair on either side of the ligation junction dramatically decreases the efficiency of the enzyme to ligate two adjacent oligos annealed to a template strand. Under optimal conditions, T4 DNA ligase is capable of distinguishing substrates which are perfectly matched at the ligation site from those which contain a mismatched base pair [66]. Under these conditions, dsDNA substrates containing two or three adjacent mismatches apparently do not ligate at all [68]. The stringent requirement of complementary ends has been applied in the rapid detection of genetic diseases by the ligase chain reaction (LCR) and related techniques [69-71]. Under appropriate conditions, especially as defined by DNA concentration and by incubation temperature, the specificity of T4 DNA ligase to mediate the joining of correctly base paired cohesive end sequences is sufficient to provide the ligation fidelity necessary in DNA indexing systems.

1.3 Thesis Overview

The focus of this Ph.D. project is the demonstration of DNA indexing as an integrated, high-throughput, high-efficiency technology adaptable to various applications, and includes the development of DNA indexing strategies for directed mapping and sequencing; application of indexed genomic profiling to bacterial strain and species differentiation; and the modification of 3'-end cDNA indexing protocols to facilitate global gene expression profiling in *Saccharomyces cerevisiae*.

1.3.1 Development of DNA Indexing Strategies for Directed Mapping and Sequencing

The plasmid pUC19 was selected to form the basis of a model system for protocol development and troubleshooting of DNA indexing strategies. Optimal conditions for *Fok*I digestion of pUC19 DNA were established. The pUC19 model system was indexed using the single-primer P/P indexing strategy in order to demonstrate the targeted amplification of each of the pUC19 *Fok*I fragments, and to observe the generation of non-targeted repeated-end fragments. Ligation conditions providing an appropriate balance between ligation fidelity and ligation efficiency were established. The ability of the P/NoP indexing strategy to eliminate the production of false positives due to repeated-end amplification was demonstrated.

Attempts to provide directionality in amplifications of indexed templates through the use of two indexer sets, each with a different core sequence and common primer, produced an artifact characterized as a nontypical form of primer-dimer (PD), capable of out-competing the amplification of correctly-indexed target fragments. The Bam and BamCC indexer sets were designed to target and amplify indexable fragment classes while preventing the amplification of repeated-end fragments, to provide the means for directional cycle sequencing of amplified indexed fragments, and to eliminate PD artifact amplification from indexing PCR reactions.

A simple exercise demonstrating the utility of DNA indexing in DNA mapping and sequencing was performed. The effective use of the Bam/BamCC indexing sets to perform compound-primer P/NoP indexing was demonstrated by the efficient indexing and amplification of each of the pUC19 target fragments. Direct cycle sequencing of the amplified indexed pUC19 fragments was performed. Indexing-based directionally-sequenced pUC19 templates were aligned to an indexing-based restriction map of pUC19 constructed by jigsaw assembly. This demonstrated the successful application of the compound-primer strategy to the pUC19 model system, in a manner that permitted the complete characterization and sequencing of the system. Model system characteristics presenting challenges to indexing approaches mimicking those presented in more complex systems were investigated, and indexing strategies were developed to meet those challenges. On the basis of these results and the basic principles involved in DNA indexing approaches, a proposal for a future application of DNA indexing was developed (see **Section 5.2**).

1.3.2 Bacterial Strain and Species Differentiation by Indexed Genomic Profiling (IGP)

An indexing-based approach to microbial molecular subtyping was developed and demonstrated. Existing indexing protocols were adapted for the complexity of microbial genome analysis and to provide increased information density in experimental data. Initial application of the modified protocols to the molecular fingerprinting and differentiation of several *E. coli* strains was accompanied by predictive modeling based on the published genomic DNA sequence of *E. coli* strain MG1655. Indexed genomic profiles were generated from clinical isolates and reference strains of several *Staphylococcus* species. Indexed genomic profiling (IGP) provides excellent discriminatory power in the form of an information-dense molecular fingerprint derived by objective sampling of microbial genetic structure. IGP provides an attractive alternative to current methods for microbial typing in clinical or research laboratory environments due to its specificity, widespread applicability, reproducibility, and potential for high-throughput application.

1.3.3 Global Gene Expression Profiling of *Saccharomyces cerevisiae* by 3'-end cDNA Indexing

The purpose of this investigation was to develop modified 3'-end cDNA indexing protocols in order to facilitate global gene expression profiling in *S. cerevisiae*. Double-stranded cDNA populations were synthesized from yeast total cellular mRNA in preparation for cDNA indexing analysis. The cDNA populations were demonstrated to be representative of the mRNA populations from which they were derived. *Fok*I digestion of cDNA populations generated indexable 3'-terminal cDNA fragments predicted from ORF sequence data of the *S. cerevisiae* genome. Indexers were ligated to the complementary cohesive end sequences of targeted *Fok*I-digested 3'-terminal cDNA fragments within a complex cDNA restriction digest. The indexed fragments were amplified using an indexing primer and transcript-specific primers. The selective capture of target fragments ligated to biotinylated indexers by streptavidin-coated paramagnetic beads was employed to reduce amplification reaction complexity, following the determination of stringent wash regimens eliminating nonspecific binding of nonbiotinylated cDNA. Using a series of artificial poly(A)-tailed indexable constructs, anchored GC-rich poly(T) primers were found to provide improved priming efficiency for cDNA population synthesis and for amplification relative to several other anchored poly(T) primer conformations. Artificial poly(A)-tailed indexable constructs were also used to evaluate the efficiency of 3'-terminal cDNA fragment amplification from templates bound to paramagnetic beads, to determine the amount of particular transcript species needed for target amplification following the ligation of NoP indexers, and to determine improved PCR cycling parameters. Ligation conditions providing high ligation fidelity with *Taq* DNA ligase were incorporated into the modified cDNA indexing protocol set.

Using the modified cDNA indexing protocols, differential gene expression profiles for yeast cultures exposed to various environmental stimuli were identified. Expression of the *GAL1* transcript was observed in yeast grown in galactose-containing medium, while no *GAL1* expression was detected in yeast grown in glucose-containing medium. Increased expression of the *BOP3* transcript was observed in cDNA populations derived from pheromone-treated yeast cultures relative to cDNA populations derived from untreated cultures. These findings were well-correlated with published data obtained by established methods of gene expression analysis [72, 73]. A limited survey of gene expression changes in yeast responding to saline shock performed using a small number of indexers generated results compatible with published data obtained in studies of yeast salt shock response using cDNA microarrays.

Analysis of 3'-end cDNA indexing data by automated fluorescence-based DNA sequencing instrumentation revealed the reproducibility of cDNA indexing profiles generated from independent parallel indexing ligations targeting individual cDNA populations and from distinct cDNA populations derived from parallel yeast cultures grown under identical conditions. Distinct indexed gene expression profiles were generated from cDNA populations derived from yeast cultures grown in the presence of differing environmental stimuli. Differences in the level of amplification of specific indexed 3'-terminal cDNA fragments were observed, indicating differences in the level of expression of specific mRNA transcripts between saline-treated and untreated yeast cultures. However, 3'-end cDNA indexing data sets were poorly correlated with data from published studies of saline shock response in *S. cerevisiae*. Unanticipated fragments were amplified, and certain anticipated indexed 3'-end cDNA fragments were not detected in 3'-end cDNA indexing data sets, indicating that refinement of the 3'-end cDNA indexing technique is necessary for effective application to global gene expression profiling in *S. cerevisiae*.

2 Chapter II: Development of DNA Indexing Strategies for Directed Mapping and Sequencing

2.1 INTRODUCTION

The plasmid pUC19 was selected to form the basis of a model system for protocol development and troubleshooting of DNA indexing strategies. Optimal conditions for *Fok*I digestion of pUC19 DNA were established. The pUC19 model system was indexed using the single-primer P/P indexing strategy in order to demonstrate the targeted amplification of each of the pUC19 *Fok*I fragments, and to observe the generation of non-targeted repeated-end fragments. Ligation conditions providing an appropriate balance between ligation fidelity and ligation efficiency were established. The ability of the P/NoP indexing strategy to eliminate the production of false positives due to repeated-end amplification was demonstrated.

Attempts to provide directionality in amplifications of indexed templates through the use of two indexer sets, each with a different core sequence and common primer, produced an artifact characterized as a nontypical form of primer-dimer (PD), capable of out-competing the amplification of correctly-indexed target fragments. The Bam and BamCC indexer sets were designed to target and amplify indexable fragment classes while preventing the amplification of repeated-end fragments, provide the means for directional cycle sequencing of amplified indexed fragments, and eliminate PD artifact amplification from indexing PCR reactions.

A simple exercise demonstrating the utility of DNA indexing in DNA mapping and sequencing was performed (outlined in FIGURE 2.1). The effective use of the Bam/BamCC indexing sets to perform compound-primer P/NoP indexing was demonstrated by the efficient indexing and amplification of each of the pUC19 target fragments. Direct cycle sequencing of the amplified indexed pUC19 fragments was performed. Alignment of indexing-based directionally-sequenced pUC19 templates to an indexing-based restriction map of pUC19 constructed by jigsaw assembly. This demonstrated the successful application of the compound-primer strategy to the pUC19

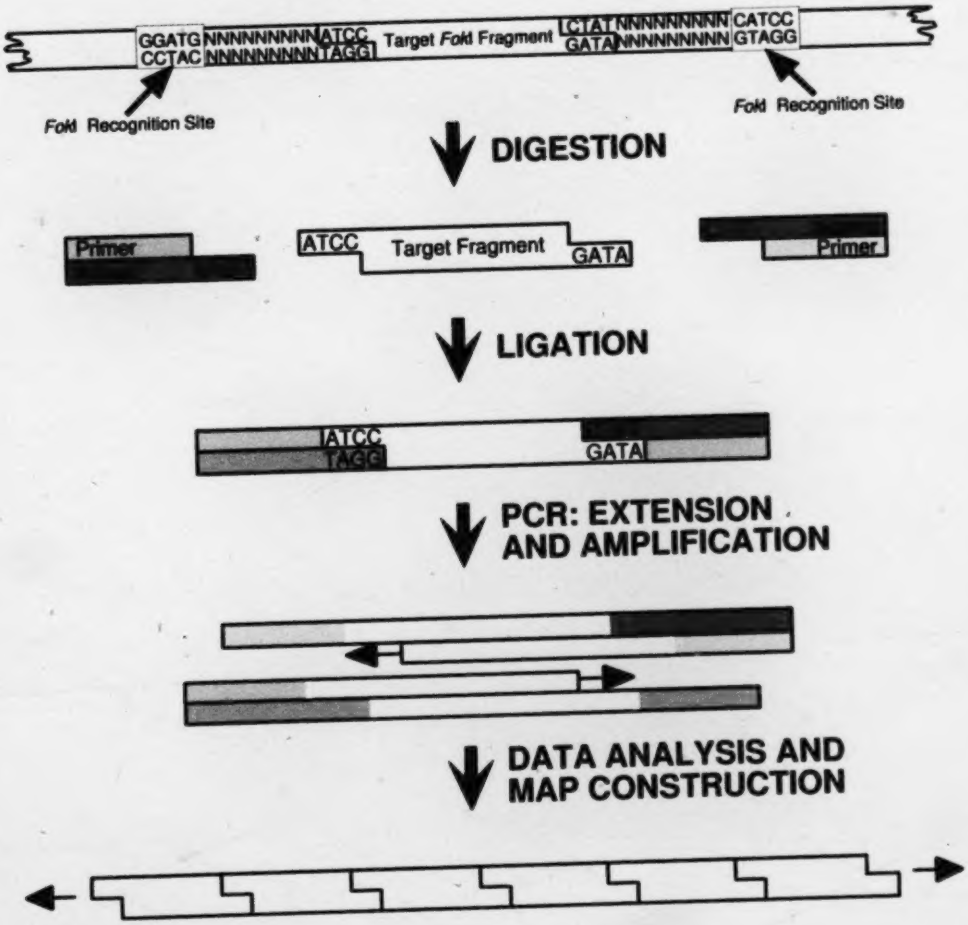

GGATG NNNNNNNNN ATCC
CCTAC NNNNNNNNNN TAGG
Target FokI Fragment
CTAT NNNNNNNNN CATCC
GATA NNNNNNNNNN GTAGG
FokI Recognition Site
FokI Recognition Site
DIGESTION
Primer
ATCC
Target Fragment
GATA
Primer
LIGATION
ATCC
TAGG
GATA
PCR: EXTENSION AND AMPLIFICATION
DATA ANALYSIS AND MAP CONSTRUCTION

FIGURE 2.1: Schematic flow diagram of DNA indexing.

model system, in a manner that permitted the complete characterization and sequencing of the system. Model system characteristics presenting challenges to indexing approaches mimicking those presented in more complex systems were investigated, and indexing strategies developed to meet those challenges. On the basis of these results and the basic principles involved in DNA indexing approaches, a proposal for a future application of DNA indexing was developed (see **Section 5.2**).

2.1.1 Selection of pUC19 as a model system for DNA indexing

Several viral and plasmid DNAs (including adenovirus-2, bacteriophage λ, simian virus-40, plasmid pBR322, and plasmid pUC19) were considered as candidates to form the basis of a model system for protocol development and troubleshooting of DNA indexing strategies. Candidate DNA sequences were evaluated according to criteria including restriction fragment number, restriction fragment size, cohesive end sequence and the ability to simulate the indexing characteristics of more complex DNA target systems. On the basis of these criteria, the small (2686 bp) *E. coli* cloning vector plasmid pUC19, constructed by Yanisch-Perron *et al.* [74] from regions of plasmid pBR322 and the M13mp19 cloning vector, was selected to be the DNA indexing model system.

There are five *Fok*I recognition sites in pUC19, producing five restriction fragments varying in size from 185 bp to 1335 bp (FIGURE 2.2; see also TABLE 2.1). This size range is representative of the fragment sizes which indexing approaches must manipulate in more complex DNAs. The small number of fragment types present, combined with the fully-characterized nature of the plasmid, was expected to facilitate the analysis and correction of false-positive generation due to misligation or multiple-event ligation. Contrary to statistical expectations for a random DNA sequence of comparable length, two of the five *Fok*I sites in pUC19 generate identical cohesive-end sequences. This characteristic allows the pUC19 model system to be used in the development of indexing strategies capable of resolving artifactual fragment amplification due to *repeated-end fragments*, (discussed in **Section 2.3.4**) as anticipated in complex systems. The selection of pUC19 as a model system for DNA

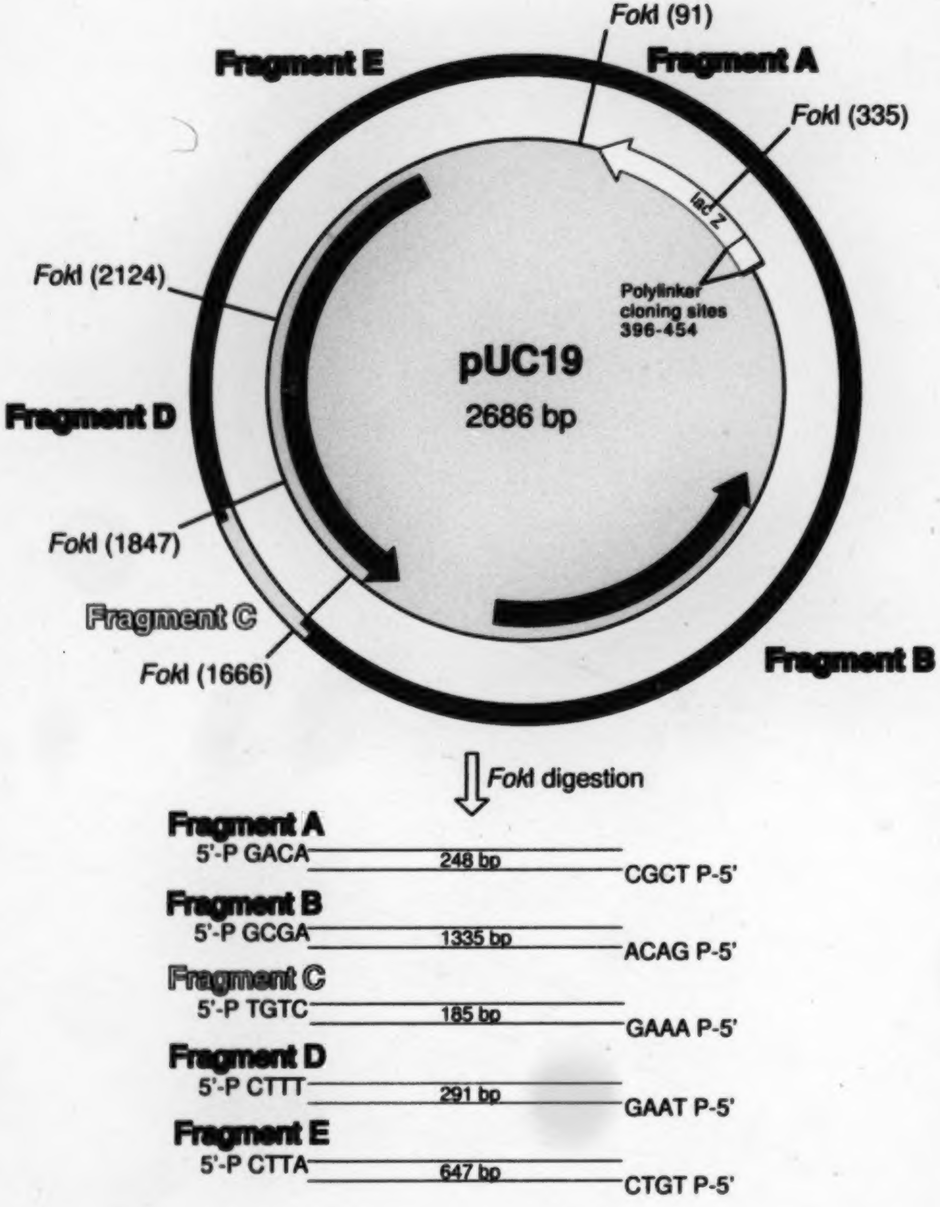

FokI (91)
Fragment E
Fragment A
FokI (335)
lac Z
Polylinker cloning sites 396-454
FokI (2124)
pUC19
2686 bp
Fragment D
FokI (1847)
Fragment C
FokI (1666)
Fragment B
FokI digestion
Fragment A
5'-P GACA
248 bp
CGCT P-5'
Fragment B
5'-P GCGA
1335 bp
ACAG P-5'
Fragment C
5'-P TGTC
185 bp
GAAA P-5'
Fragment D
5'-P CTTT
291 bp
GAAT P-5'
Fragment E
5'-P CTTA
647 bp
CTGT P-5'

FIGURE 2.2: ***FokI* restriction map of plasmid pUC19.**

TABLE 2.I: pUC19 target fragment end sequences and fragment sizes.

pUC19 Fragment	Fragment End Sequences		Indexer End Sequences		Fragment Size	Indexed Fragment Size	
	Left	Right	Left	Right		P/P Indexers	P/NoP Indexers
A	GACA	TCGC	TGTC	GCGA	248	288	290
B	GCGA	GACA	TCGC	TGTC	1335	1375	1377
C	TGTC	AAAG	GACA	CTTT	185	225	227
D	CTTT	TAAG	AAAG	CTTA	291	331	333
E	CTTA	TGTC	TAAG	GACA	647	689	691
AB	GACA	GACA	TGTC	TGTC	1579	1619	-
CDE	TGTC	TGTC	GACA	GACA	1115	1155	-

indexing therefore permitted the testing and troubleshooting of indexing protocols on a simple system that presented challenges to indexing techniques analogous to those presented by systems of greater complexity.

*Fok*I exists as a monomer and recognizes an asymmetric DNA sequence 5'-GGATG-3', cleaving DNA phosphodiester groups 9 bp and 13 bp away (in the 5' direction) from the recognition site. It is the most extensively characterized of the Type IIS restriction endonucleases, and acts as an archetypal model for other members of the subclass [75-82]. These characteristics made *Fok*I the most attractive choice of enzyme to form the basis for indexing model system development and application.

2.1.2 Selection of T4 DNA ligase for development of the pUC19 indexing model system

T4 DNA ligase was the DNA joining enzyme employed in investigations of the pUC19 model system. As previously discussed, the specificity of T4 DNA ligase to mediate the joining of correctly base paired cohesive end sequences is sufficient to provide the ligation fidelity necessary in DNA indexing systems when employed under appropriate conditions. Defining those appropriate conditions represents an important step in the development of a DNA indexing model system.

2.1.3 Selection of DNA polymerases for indexing applications

2.1.3.1 Selection of Taq *DNA polymerase for development of the pUC19 indexing model system*

Despite the range of other DNA polymerases commercially available for use in PCR, *Taq* DNA polymerase was selected for use in the development of DNA indexing protocols using the pUC19 model system. Evaluations of *Taq* DNA polymerase in the amplification of indexing reactions by Unrau and Deugau [22] demonstrated this enzyme's amplification of indexed DNA fragments of similar size to those of the pUC19 system to be adequate for the purposes of protocol development. Widespread use of *Taq* DNA polymerase in the development of other PCR-based technologies [83-90] indicated that this enzyme would provide an effective tool for the development of indexing protocols that, if required, could be easily modified to employ alternative enzymes with more task-specific features. *Taq* DNA polymerase has the advantage of being a well-characterized enzyme [91-97] with clearly-established functional characteristics in a variety of DNA amplification systems [98-104]. Basic PCR protocols for the development of the pUC19 DNA indexing model system were adapted from Unrau and Deugau [22] with minor modifications, as described in **Section 2.2.6**.

2.1.3.2 Selection of PfuTurbo™ *DNA polymerase for amplification of indexed sequencing templates*

For amplification of indexed pUC19 fragments in preparation for their use as templates for cycle sequencing, *PfuTurbo*™ DNA polymerase (Stratagene Cloning Systems, La Jolla CA) was employed. A blend of cloned *Pfu* DNA polymerase and a betaine-like factor that improved product yields without altering DNA replication fidelity, *PfuTurbo*™ DNA polymerase amplifies complex targets in higher yield and with greater fidelity than *Taq* DNA polymerase (data not shown; also references [105-107]). *PfuTurbo*™ DNA polymerase provided an attractive alternative to *Taq* DNA polymerase for the amplification of indexed pUC19 fragments in preparation for their use as templates for cycle sequencing.

2.2 MATERIALS AND METHODS

2.2.1 Evaluation and selection of DNA indexing model system

Several viral and plasmid DNAs (including adenovirus-2, λ phage, simian virus-40, pBR322, and pUC19) were considered as candidates to form the basis of a DNA indexing model system. The sequences of the candidate DNAs were downloaded from GenBank via Entrez (National Center for Biotechnology Information, Bethesda MD) and their *Fok*I restriction endonuclease cleavage patterns were analyzed with DNA Strider v1.2 (Dr. C. Marck, Gif-sur-Yvette, France). pUC19 (2686 bp; GenBank Accession #X02514) was selected as the most suitable DNA indexing model system among the candidate DNA sequences, based on criteria including restriction fragment number, restriction fragment size, cohesive end sequence and ability to simulate the indexing characteristics of more complex DNA target systems.

2.2.2 Digestion of pUC19 DNA with *Fok*I restriction endonuclease

*Fok*I digests of pUC19 were performed in NEBuffer 4 (50 mM KOAc; 20 mM Tris-acetate (pH 7.9 @ 37°C); 10 mM $Mg(OAc)_2$; and 1 mM DTT)(New England BioLabs, Beverly MA). One microgram of pUC19 plasmid DNA was typically incubated with 1 U *Fok*I restriction endonuclease (New England BioLabs) in 1x NEBuffer 4 in a reaction volume of 20 µl for 1 h at 20°C. Digestion was halted by heat denaturation of the restriction enzyme at 65°C for 20 min. Early digests were purified by extraction with 10 µl (1/2 volume) chloroform-isoamyl alcohol (24:1)(CIAA), followed by ethanol precipitation and re-dissolution of the digested DNA in 20 µl Tris-EDTA (10 mM Tris, 1 mM EDTA, pH 7.6)(TE). Later digests were not chloroform-extracted, because this step was found to be superfluous. Digest quality was evaluated by electrophoresing a 1-µl aliquot of the digest (containing 50 ng, or about 30 fmol, of pUC19 DNA), and 50 ng of undigested pUC19 DNA, on a 1.8% agarose gel at 100 V for 50 min. The digest was diluted to a working stock concentration of 1 ng/µl in TE and was stored at -20°C until required.

2.2.3 Synthesis of indexing oligonucleotides

The α and γ indexing oligonucleotide (oligo) sets were synthesized by the Molecular Biology Service Unit of the Department of Biological Sciences (University of Alberta, Edmonton AB) using an Applied Biosystems Model 392 RNA/DNA Synthesizer (Applied Biosystems, Foster City CA). The phosphorylated Bam indexing oligos, Bam primers and BamCC primers were synthesized by the University Core DNA & Protein Services Oligonucleotide Synthesis Laboratory (University of Calgary, Calgary AB) on a similar instrument. The nonphosphorylated BamCC indexing oligo mixes were donated by Dr. H. Heyneker and were synthesized using proprietary 96-well massively-parallel oligo synthesis technology (ProtoGene

Laboratories, Menlo Park CA). All oligos from all sources were received in lyophilized form.

The lyophilized oligos were re-dissolved in an appropriate volume of TE, based on the supplier's analysis of the quantity of oligo provided. In addition, quantitation of the resuspended oligo was performed by spectrophotometric determination of the solution's UV absorbance, using a Model 8450A Diode Array Spectrophotometer (Hewlett Packard, Palo Alto CA). From this data, 200 pmol/µl storage stocks of each oligo were prepared. All storage stocks were stored at -20°C until required. A working stock of each single-stranded indexing oligo was prepared from an aliquot of the storage stock by diluting it to 20 pmol/µl in TE. (In the case of the non-phosphorylated BamCC-series indexing oligo mixes provided by ProtoGene, the storage stocks were diluted to a concentration of 20 pmol/µl of total oligo. This dilution reflects a concentration of 5 pmol/µl for each of the 4 indexing-strand oligos present in a particular BamCC-series oligo mix.) All handling and pipetting of indexing oligos and all subsequent manipulations of indexing reagents and reactions were performed with disposable self-sealing ART Barrier aerosol-resistant pipette tips (Molecular BioProducts, San Diego CA) to prevent aerosol contamination of indexing oligos and PCR reactions. See TABLE 2.2 for a list of indexing oligos used in this investigation.

2.2.4 Annealing of indexing oligonucleotides

Equimolar quantities of a primer oligo and an indexing-strand oligo were annealed to form a functional double-stranded indexer. One microlitre of 20 pmol/µl working stock of the appropriate primer and 1 µl of 20 pmol/µl indexing oligo working stock were thoroughly mixed with 8 µl TE. The 10 µl annealing reaction volume had a concentration of 2 µM of each oligo. The reaction was heated to 65°C for 5 min, and then cooled at the rate of 1°C/10s until the reaction's temperature was 16°C. Three hundred and ninety microlitres of TE were added to the annealing reaction, bringing

TABLE 2.2: Indexing oligo sequences employed in DNA indexing investigations.

Indexing Oligo	Sequence	Length (nt)	T_m (°C)	Purpose
α Indexing Strand	5'- *P* -**NNNN**TTGTCTTCGCATCCTGTACC -3'	24	see primer	Single-primer P/P indexing,
α Primer Strand	5'- GGTACAGGATGCGAAGACAA-3'	20	54.8	double-primer P/NoP indexing
γ Indexing Strand	5'- *OH* -**NNNN**CTATCGATGCATGCTGTACC -3'	24	see primer	Double-primer P/NoP indexing
γ Primer Strand	5' -GGTACAGCATGCATCGATAG-3'	20	54.8	
Bam Indexing Strand	5'- *P* -**NNNN**TTGTCTTCGGATCCTGTACC -3'	24	see primer	Directional single-primer P/NoP indexing
Bam Primer	5'- GGTACAGGATCCGAAGACAA-3'	20	54.8	
Bam Indexing Strand	5'- *P* -**NNNN**GGTTGTCTTCGGATCCTGTACC-3'	24	see primer	
BamCC Primer	5' -GGTACAGGATCCGAAGACAACC-3'	22	58.2	
ACT1 Forward	5'- ATCCAAGCCGTTTTGTCCTTGTA -3'	23	58.6	Yeast cDNA Test Fragment amplification
ACT1 Reverse	5'- ATGGACCACTTTCGTCGTATTC -3'	22	56.1	
FUS1 Forward	5'- CACGCCAGATTCACAAATCA -3'	20	54.6	
FUS1 Reverse	5'- CAGTCGTATTCTTGGAGACAGTCA-3'	24	57.6	
STE12 Forward	5'- GGATTTTGATGAATCTCGGC-3'	20	52.6	
STE12 Reverse	5'- GGCATCTGGAAGGTTTTTATCGG -3'	23	57.6	
STE2 Forward	5'- CCACAATTTTACTTGCATCCTC -3'	22	53.5	
STE2 Reverse	5'- TACATGTCGACGGGTTCAACTT-3'	22	57.8	
Poly(T)$_{18}$-A	5'- TTTTTTTTTTTTTTTTTTA-3'	19	36.6	Anchored poly(T)$_{18}$ priming of cDNA synthesis
Poly(T)$_{18}$-C	5'- TTTTTTTTTTTTTTTTTTC -3'	19	38.1	
Poly(T)$_{18}$-G	5'- TTTTTTTTTTTTTTTTTTG-3'	19	38.5	
Poly(T)$_{35}$-A	5'- TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTA-3'	36	51.0	Anchored poly(T)$_{35}$ priming of cDNA synthesis
Poly(T)$_{35}$-C	5'- TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTC -3'	36	51.7	
Poly(T)$_{35}$-G	5'- TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTG -3'	36	51.8	
BamPolyTPrimer	5'- GGTACAGGATCCGAAGACAA(T)$_{35}$-3'	55	62.5	Construction of FakePolyT Fragment
PolyA-TGTCindexer	5'- TGTC(A)$_{35}$ -TTGTCTTCGGATCCTGTACC -3'	59	64.0	for cDNA indexing protocol development
GCRichPoly(T)$_{16}$-A	5'- GGGCACGCTTTTTTTTTTTTTTTTA-3'	25	54.8	G/C-rich anchored poly(T)$_{16}$ priming of cDNA synthesis and 3'-end cDNA indexing
GCRichPoly(T)$_{16}$-C	5'- GGGCACGCTTTTTTTTTTTTTTTTC-3'	25	55.8	
GCRichPoly(T)$_{16}$-G	5'- GGGCACGCTTTTTTTTTTTTTTTTG-3'	25	56.0	

the concentration of annealed indexer to 50 nM. The 50 fmol/µl indexer stock was stored at 4°C.

In the case of the BamCC oligo mixes, 4 µl of 20 pmol/µl BamCC primer stock and 4 µl of 20 pmol/µl [total oligo] BamCC nonphosphorylated indexing oligo mix were added to 2 µl TE and mixed thoroughly. The 10 µl annealing reaction volume contained 8 µM BamCC primer, and 2 µM of the four BamCC indexing oligos, each with a specific cohesive end sequence. The reaction was heated to 65°C for 5 min, and then cooled at the rate of 1°C/10s until the reaction's temperature was 16°C. Three hundred and ninety microlitres of TE were added, diluting the concentration of each of the four annealed BamCC indexers to 50 nM. The 50 fmol/µl/indexer (200 fmol/µl total indexer) stock was stored at 4°C.

2.2.5 Ligation of indexers to pUC19 DNA: standard conditions

A typical pUC19 indexing ligation reaction contained 1 ng (~0.6 fmol) of *Fok*I-digested pUC19 DNA, 50 fmol of Indexer 1 and 50 fmol of Indexer 2. Indexer 1 was a phosphorylated indexer bearing the cohesive end sequence complementary to that of one of the cohesive ends generated by *Fok*I digestion of the pUC19 target fragment. Indexer 2 was typically a phosphorylated or non-phosphorylated indexer bearing the cohesive end sequence complementary to the other cohesive end of the pUC19 target fragment. Alternately, Indexer 2 was a mix of four non-phosphorylated BamCC indexers, one bearing the cohesive end sequence complementary to the second cohesive end of the target fragment and the other three indexers differing from that sequence at the fourth base position from the 5'-end of the indexing strand. The ligation reaction volume was 19 µl prior to the addition of ligase, and contained 50 mM Tris-HCl (pH 7.8), 10 mM $MgCl_2$, 10 mM DTT, 1 mM ATP, and 10 ng bovine serum albumin (BSA). A positive ligation control containing indexers and DNA previously demonstrated to be ligated successfully was added as an extra sample in each experiment set. The ligation reaction was warmed to 40°C for two minutes to prevent premature annealing of indexers to target fragment cohesive ends, and brought to 37°C.

The temperature of the reaction was allowed to equilibrate at 37°C for one minute. Forty cohesive-end ligation units (U_L) (1 U_L = 0.015 Weiss units) of T4 DNA ligase (New England BioLabs) were added, the reaction was mixed thoroughly by pipetting up and down, and the reaction was incubated at 37°C for 1 h. The reaction was terminated by heat inactivation of the DNA ligase at 65°C for 20 min. The terminated ligation reaction was then stored at 4°C.

2.2.6 Amplification of indexed pUC19 fragments by Polymerase Chain Reaction (PCR) : standard conditions

pUC19 target fragments to which indexers had been ligated were amplified by the polymerase chain reaction (PCR). PCR reactions were assembled in 200-µl thin-walled PCR tube. Two microlitres of a pUC19 ligation reaction were added to a PCR reaction mix containing 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 1.5 mM $MgCl_2$, 200 µM dNTPs, and 20 pmol of each indexing primer needed to amplify the target fragment. If the same priming sequence was present on both ends of the target fragment, then 40 pmol of the single primer required to amplify the fragment were added. In addition to the ligated indexing samples to be amplified, a negative PCR control containing all reagents and primers but no ligated DNA and a positive PCR control containing an aliquot of a ligation stock previously demonstrated to amplify successfully were used in each experiment.

Early PCRs were assembled to a volume of 49.5 µl prior to the addition of heat-stable DNA polymerase. The tubes were heated to 95°C for 2 min, cooled to 80°C prior to the addition of 0.5 µl (1.25 U) *Taq* DNA polymerase (Gibco BRL Life Technologies, Rockville MD), and then re-heated to 95°C to begin amplification. This "hot start PCR" approach was found to be unnecessary when a heat-stable DNA polymerase complexed with a heat-labile antibody was employed (e.g. Platinum™ *Taq* DNA Polymerase, Gibco BRL), and later amplification reactions were assembled with the DNA polymerase included prior to heating.

Thermal cycling of PCR reactions was performed on an PTC-100 Peltier Thermal Cycler (MJ Research, Waltham MA). PCR reactions subjected to a "hot start" were heated to 95°C for 2 min, cooled to 80°C for the addition of heat-stable polymerase, and re-heated to 95°C for 1 min. If no "hot start" was used, the temperature of the reaction was brought to 95°C for 1 min to dissociate the fragment strands (and also any improperly annealed primers). The amplification program "NEW55-62" was typically used to amplify indexed pUC19 target fragments. In the first cycle of this program, the amplification reactions were heated to 95°C for 1 min for the melting step; cooled to 55°C for 1 min for the annealing step; and held at 72°C for 2 min during the elongation step. The annealing step of each subsequent cycle was cooled only to 62°C, reducing the likelihood of non-specific priming or "mis-priming" events in those cycles and reducing the background artifact level in the amplification. Following a typical amplification regimen of 30 cycles, the PCR reaction was held at 72°C for 10 min to permit elongation of any partially-extended templates, and the reaction was held at 4°C.

2.2.7 Agarose gel electrophoresis of amplified indexed pUC19 fragments

Following amplification, indexing reactions were analyzed by agarose gel electrophoresis. A 1.8% (w/v) agarose gel containing 200 µg/ml ethidium bromide (EtBr) was prepared in Tris-acetate•EDTA (40 mM Tris-acetate, 1 mM EDTA, pH 8.0; TAE) and set in a gel form with an appropriate lane comb. The gel was submerged in TAE in the buffer reservoir of a Sub Cell electrophoresis apparatus (BioRad, Hercules CA). Two microlitres of loading dye (0.25% bromophenol blue, 40% sucrose) was added to 10 µl of each PCR reaction, mixed thoroughly, and loaded into one well of the gel. A DNA molecular size standard was prepared by mixing 2 µl of loading dye and 8 µl of ddH_2O with 1-2 µg of DNA size standard stock. After all samples and standards were loaded onto the gel, electrophoresis was performed at a field strength of 6 V/cm for 50 minutes.

2.2.8 UV transillumination of ethidium bromide-stained agarose gels

After electrophoresis of the indexing DNA samples was complete, the gel was removed from the buffer reservoir and visualized on a 312-nm UV transilluminator

(FisherBiotech Electrophoresis Systems FBTI 816, Fisher Scientific, Pittsburgh PA). Gels analyzed early in the investigation were photographed with a handheld FisherBiotech Photodocumentation Camera (Fisher Scientific) on high-speed Polaroid 667 black-and-white film (Polaroid Corporation, Cambridge MA). Later gels were recorded with a Pixera PVC 100C high-resolution full-colour digital camera (Pixera Corporation, Los Gatos CA) mounted on an Ultra-Lum transilluminator hood chassis. The data were analyzed via a suite of software for the Macintosh PowerPC computer including Pixera Studio VCS v1.2 (Pixera Corporation), NIH Image v1.6 (National Institutes of Health, Bethesda MD) and RFLP v1.2 (Advanced American Biotechnology, Fullerton CA).

2.2.9 Protocols relating to the study of primer-dimer

2.2.9.1 Isolation of primer-dimer (PD) DNA from agarose gel by syringe extraction

Amplified primer-dimer was isolated from agarose gel electophoresis using a syringe-extraction protocol modified from Li and Ownby, 1993 [108]. A set of indexing PCR reactions with constituents demonstrated to produce primer-dimer was assembled in parallel and amplified according to the indexing PCR conditions previously described. The completed reactions were pooled. The total pooled reaction volume was loaded into multiple lanes of a single 1.8% agarose preparative gel and electrophoresed as described. Primer-dimer bands on the gel were visualized by UV transillumination. A razor blade was used to dissect the ethidium-bromide-visualized bands of PD DNA from the gel in slices with average size 5 x 2 x 1.5 mm. The nozzle of a 1 cc slip-tip syringe (Becton Dickinson, Franklin NJ) was pared down to reduce the syringe's dead volume, and the syringe's plunger was removed. A small disc of Grade 1 Chromatography paper (Whatman International Ltd, Maidstone, Kent UK) with a diameter slightly larger than the syringe's inner diameter was inserted into the syringe and pressed flat across the bottom opening of the syringe using the plunger. The syringe plunger was again removed. The agarose slices were placed into the syringe, and the plunger re-inserted, pushing the slices to the bottom of the syringe onto the filter. Pressure was applied to the plunger, forcing buffer out of the gel through the filter and leaving a dense agarose slurry residue. Effluent buffer containing PD DNA was captured in a 500-µl Eppendorf microcentrifuge tube (Eppendorf AG, Hamburg, Germany) as it passed through the syringe. The amount of PD DNA in the effluent of a single slice was estimated to be at least 2 ng/µl.

Following extraction of PD DNA from preparative agarose gel slices, the DNA was ethanol-precipitated to concentrate the sample and to remove salts, ethidium bromide and other impurities. Two volumes of cold (-20°C) 95% ethanol were added to the effluent, and the DNA was precipitated overnight at -20°C. The sample was centrifuged at 13,200 rpm for 15 min at room temperature in an IEC MicroMax microcentrifuge (Fisher Scientific). The supernatant was decanted from the microcentrifuge tube, leaving a visible pellet of precipitated DNA on the tube wall. The pellet was washed with 100 µl of cold 70% ethanol, the sample tube was re-centrifuged for 2 min at 13,200 rpm and the supernatant was again decanted. The open tube was stored at room temperature until all liquid had evaporated. The DNA pellet was re-dissolved in 20 µl TE overnight at 4°C.

2.2.9.2 Preparation of primer-dimer amplification product for cloning

To prepare the PD amplification product for blunt-end cloning into a pUC19 sequencing vector, the insert was "polished" by extending any incompletely-replicated DNA fragments with the Klenow fragment of *Escherichia coli* DNA polymerase I. One microlitre of 2.5 mM dNTPs, 0.6 µl of 5.8 U/µl Klenow fragment (Gibco BRL), and 2.4 µl of 10x REact 2 buffer (final buffer concentration: 10 mM $MgCl_2$, 50 mM NaCl, 50 mM Tris-HCl, pH 8.0; Gibco BRL) were added to the 20-µl volume of re-dissolved DNA for a total reaction volume of 24 µl. The polishing reaction was incubated at room temperature for 15 min, followed by heat denaturation of the enzyme at 75°C for 15 min.

The polished PCR products were then 5'-phosphorylated with T4 polynucleotide kinase. Twenty units of kinase (10 U/µl; Gibco BRL) were added to the reaction following the addition of 3.1 µl of polynucleotide kinase (Forward) reaction buffer (final buffer concentration: 15 mM $MgCl_2$, 50 mM NaCl, 25 mM KCl, 500 µM 2-mercaptoethanol, 85 mM Tris-HCl, pH 8.0; Gibco BRL) and 1.5 µl of 10 mM ATP. The kinase reaction was incubated at 37°C for 1 h, and the kinase heat-denatured at 65°C for 15 min.

The prepared PD DNA sample was purified of enzymes, salts and nucleotides with a QIAquick Nucleotide Removal Kit (QIAgen Corp.). The manufacturer's suggested protocol was followed, permitting recovery of double-stranded oligonucleotides larger than 17 bp in length. The PD DNA insert (whose length prior to sequencing was estimated to be 30-38 bp) was eluted from the QIAquick column in a final volume of 30 µl ddH_2O. Concentration of the eluted DNA sample was determined by UV spectrophotometry to be 3 ng/µl.

2.2.9.3 Blunt-end cloning of PD DNA into pUC19 sequencing vector

The sequencing vector pUC19 was cleaved with *Hinc*II restriction endonuclease (recognition sequence GTPy/PuAC) to provide a blunt-ended insertion site for PD DNA in preparation for sequencing with the M13mp18 sequencing primer set. Three micrograms of pUC19 DNA was digested with 10 U *Hinc*II (New England BioLabs) in a digest volume of 20 µl containing 100 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT, 50 mM Tris-HCl, pH 7.9 @ 25 °C (NEBuffer 3; New England BioLabs) and 100 ng/µl BSA. The digest was incubated at 37°C for 1 hr, and the reaction was terminated by heat denaturation at 65°C for 20 min.

Digest quality was evaluated by comparison of the *Hinc*II-digested vector with undigested pUC19 DNA by agarose gel electrophoresis. Five hundred nanograms of undigested pUC19 DNA was loaded into one well of 1.8% agarose gel, and an equal amount of the pUC19 *Hinc*II digest was loaded into an adjacent well. Electrophoresis and visualization were performed in the manner previously described. Efficient cleavage of the digested pUC19 DNA was demonstrated.

Prepared PD DNA was ligated into the *Hinc*II-digested pUC19 sequencing vector in blunt-end ligation buffer (10% w/v polyethylene glycol (PEG) 8000, 10 mM

$MgCl_2$, 10 mM DTT, 1 mM ATP, 25 ng BSA, 50 mM Tris-HCl, pH 7.8). (PEG 8000 was used as a macromolecular crowding agent that promotes the ligation of blunt ends.) Twelve hundred units of T4 DNA ligase (New England BioLabs) was used to ligate 3 pmol of PD DNA insert into 80 fmol of *Hinc*II-digested pUC19 in a total reaction volume of 55 μl. The blunt-end ligation reaction was incubated at 16°C overnight.

2.2.9.4 Transformation of competent E. coli DH5α with vector bearing PD insert

Competent *E. coli* DH5α were prepared through the use of standard molecular biology techniques [49] and stored at -80°C. When required, the tubes of frozen competent cells were thawed on ice. All plasticware and reagents were chilled to at least 4°C to prevent thermal shock to the competent cells. One hundred microlitres of thawed competent DH5α freezer stock were pipetted gently into each of three chilled thin-walled 200-μl PCR tubes, as well as into a negative control tube and a positive control tube to which 1 μg of undigested pUC19 was added. Twenty microlitres of each ligated vector were added to and gently mixed with the cells in each of the transformation reaction tubes. The transformation cultures were kept on ice for 1 hr, abruptly brought to 42°C for 90 sec, and placed in an ice bath for 5 min. The contents of each tube were gently added to 800 μl of chilled Luria Broth (LB) liquid media in a 10-ml Falcon tube. The five Falcon tubes were placed in a roller wheel at 37°C for 45 min.

Luria Broth (LB) media plates containing 100 μg/ml ampicillin, 750 μg X-gal and 1 mg IPTG were prepared following standard molecular biology techniques [49]. These Amp-LB plates were used for the growth of *E. coli* DH5α colonies transformed by pUC19 vector containing the PD insert and their identification by standard blue/white selection. Following incubation of the transformation cultures at 37°C, 200-μl aliquots of each sample were plated on Amp-LB plates (3 plates/sample). Additionally, the negative and positive control samples were plated on LB plates that did not contain ampicillin as a control for the viability of the DH5α cells. All plates were incubated overnight at 37°C.

2.2.9.5 Amplification of insert-bearing pUC19 fragments directly from transformed E. coli *DH5α colonies by AB PCR*

In addition to blue/white colony selection, colonies were screened for appropriate inserts using a modified version of Dynabeads Template Preparation Starter Kit protocol (Dynal, Great Neck NY) referred to as AB PCR [109]. Primers complementary to sequences flanking the multiple cloning site (MCS) of pUC19 [110] were used to amplify across the insert to establish that a single copy of the PD DNA insert was present in the vector. Further, the PCR product amplified from the vector in this manner provided a template for cycle sequencing across the PD insert.

The A1 oligo sequence was 5'-biotin-GCTTCCGGCTCGTATGTTGTGTG-3', and the B2 primer sequence was 5'-biotin-AAGGGGGATGTGCTGCAAGGCG-3' (University Core Oligo Synthesis Laboratory). The A1 priming site was located at position 509 to 531 bp of pUC19. The B2 priming site was located at position 315 to 337 bp. The expected length of PCR product amplified from pUC19 vector without an insert was 217 bp.

Several transformed, insert-bearing colonies were analyzed by AB PCR. In each case, a single white colony was selected from an Amp-LB plate by touching the colony with a sterile disposable 200-μl pipet tip. The colony was suspended in 20 μl of ddH_2O and the cells lysed by boiling. Cellular debris was pelleted in a microcentrifuge for 1 min at 13 200 rpm and the supernatant removed by pipetting. An AB PCR reaction was prepared containing 10 μl supernatant, 5 pmol each of primer A1 and B2, 200 μM dNTPs, 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 1.5 mM $MgCl_2$, and 2.5 U *Taq* DNA polymerase in a total reaction volume of 50 μl. The AB PCR thermal cycling program involved an initial melting step of 30 s at 96°C; 25 cycles each consisting of 30 s at 94°C, 30 s at 55°C and 1 min at 72°C; and a final extension step of 2 min at 72°C.

Four microlitres of each AB PCR reaction were electrophoresed on a 1.8% agarose gel as previously described. Amplification of recombinant pUC19 clones resulted in a PCR product roughly 255 bp in length (data not shown). AB PCR products containing an insert were cleaned with QIAquick PCR Purification Kit (QIAgen Corp.) according to the manufacturer's instructions, resuspended in TE, and quantified by UV spectrophotometry. The amplified products, whose concentrations ranged from 50 to 100 ng/µl, were used as templates for cycle sequencing of the PD insert.

2.2.9.6 Cycle sequencing of PD insert from AB PCR product

Cycle sequencing reactions were prepared using a ThermoSequenase Primer Cycle Sequencing Kit (Amersham-Pharmacia Biotech, Uppsala, Sweden) according to the manufacturer's instruction and employing fluorescently-labeled M13 forward and reverse sequencing primers (Applied Biosystems, Foster City CA). Four separate tubes were employed for a complete cycle sequencing reaction for each AB PCR template. Cycle sequencing PTC conditions involved 30 cycles of amplification, each cycle consisting of 30 s at 95°C and 30 s at 55°C. The contents of the four reaction tubes for each DNA sample were pooled.

Fifteen microlitres of streptavidin immobilized on 4% beaded agarose (Sigma-Aldrich Corporation, St. Louis MO) were added to each pooled reaction and gently mixed at room temperature for 20 min to bind the biotinylated AB PCR templates. The agarose beads were pelleted by centrifugation for 30 sec at 13,200 rpm. The supernatant was pipetted off the pellet and applied to a Microcon 30 column (Millipore Corporation, Bedford MA). The cycle sequencing samples were desalted as recommended by the manufacturer. The sample volumes recovered from the Microcon 30 columns ranged from 5 to 20 µl of desalted cycle sequencing reaction. Automated sequencing of the PD inserts was performed on an ABI Prism 377 Automated DNA Sequencer (Applied Biosystems) by the Department of Biochemistry DNA Sequencing Facility (University of Alberta).

2.2.10 Double digestion of pUC19 DNA with *Fok*I and *Sfa*NI restriction endonucleases

One microgram of pUC19 DNA (New England BioLabs) was incubated with 1 U *Fok*I restriction endonuclease (New England BioLabs) in NEBuffer 3 (100 mM NaCl; 50 mM Tris-HCl (pH 7.9 @ 25°C); 10 mM $MgCl_2$; and 1 mM DTT) in a reaction volume of 18.5 µl for 30 min at 20°C. Digestion was halted by heat denaturation of the restriction enzyme at 65°C for 20 min. An additional 1 U of *Fok*I endonuclease was added (0.5 µl of a enzyme stock diluted to 2 U/µl) and the reaction incubated for another 30 min at 20°C. The enzyme was again heat denatured at 65°C for 20 min. Two units (0.5 µl of a 4U/µl stock) of *Sfa*NI restriction endonuclease was added and the reaction was incubated for 30 min at 37°C. Following a third heat denaturation step, a second aliquot of *Sfa*NI was added to the digest and incubated for 30 min at 37°C. A final 20 min heat deactivation step at 65°C was performed. Digest quality was evaluated by electrophoresing a 1-µl aliquot of the digest and 50 ng of undigested pUC19 DNA on a 1.8% agarose gel at 100 V for 50 min. The digest was diluted to a working stock concentration of 1 ng/µl in TE and was stored at -20°C until required.

2.2.11 Amplification of indexed pUC19 fragments by *PfuTurbo*™ DNA polymerase

PfuTurbo™ DNA polymerase was employed for the amplification of indexed pUC19 fragments in preparation for their use as templates for cycle sequencing. Two microlitres of a pUC19 ligation reaction were added to a PCR reaction mix containing 0.5 µl (1.25 U) *PfuTurbo*™ DNA polymerase, 20 mM Tris-HCl (pH 8.8), 10 mM KCl, 2 mM $MgSO_4$, 10 mM $(NH_4)_2SO_4$, 0.1% Triton X-100, 100 ng/µl nuclease-free BSA, 200 µM dNTPs, and 20 pmol of each indexing primer needed to amplify the target fragment. PCR reactions were assembled to a volume of 50 µl prior to amplification using the MJ Research PTC. The amplification program "PFUPCR1" was typically used to amplify indexed pUC19 target fragments in preparation for sequencing. The temperature of the reaction was brought to 95°C for 1 min to dissociate the fragment strands (and also any improperly annealed primers). In the first cycle of this program, the amplification reactions were heated to 95°C for 1 min for the melting step; cooled to 57°C for 1 min for the annealing step; and held at 72°C for 2 min during the elongation step. The annealing step of each subsequent cycle was cooled only to 62°C, as in the case of PCR reactions employing *Taq* DNA polymerase. Following a typical amplification regimen of 30 cycles, the PCR reaction was incubated at 72°C for 10 min to permit elongation of any partially-extended templates, and the reaction was held at 4°C. Analysis of amplified fragments by agarose gel electrophoresis then proceeded in the manner previously outlined.

2.2.12 Direct cycle sequencing of amplified indexed pUC19 fragments

Amplified indexed pUC19 fragments were purified of polymerase, nucleotides and excess primers using the QIAquick PCR Purification Kit following the manufacturer's protocol. Cycle sequencing of the purified DNA fragments was performed using an ABI Prism© BigDye™ Terminator Cycle Sequencing Ready Reaction Kit according to the manufacturer's instructions. The DNA Indexing sequencing primer BamCC was used to initiate sequencing only from the end of the indexed fragments to which the non-phosphorylated BamCC indexer had been ligated. This unidirectionality ensured that the cycle sequencing reactions were not contaminated by simultaneous initiation of sequencing from both ends of the indexed template. Automated sequencing was performed by the Department of Biochemistry Sequencing Facility in a manner similar to that earlier described for the sequencing of primer-dimer inserts.

2.2.13 Mapping and assembly of contiguous pUC19 sequences

The sequence data analysis and contiguous sequence (*contig*) assembly software Sequencher v3.0 (GeneCodes Corporation, Ann Arbour MI) was used to construct a index map and a completed contig of the pUC19 vector. Automated sequence data of indexed pUC19 DNA fragments in raw chromatogram form were trimmed of indexer sequence and manually edited for obvious miscalled bases. Trimmed sequences were compared, matched and aligned by the software's contig assembly algorithm into subcontigs corresponding to each of the indexed pUC19 fragments used as templates for sequencing. Subcontigs were linked by their four-base cohesive end sequences by comparing the subcontigs to the index map. The complete contig was assembled and the results compared to the canonical GenBank sequence of the pUC19 vector.

2.3 RESULTS

2.3.1 Features of the α phosphorylated indexers and γ nonphosphorylated indexers

Indexers are comprised of two annealed synthetic oligonucleotides: an indexing strand bearing a common primer-binding sequence and a variable 4-nucleotide 5' cohesive end, and a complementary common primer strand. Selection of the appropriate indexer for each end of a targeted restriction fragment thus permits the covalent attachment of two common-primer-binding sequences onto the ends of the fragment. This permits the exclusive amplification of target fragments bearing primer-binding sequences on both ends. Four sets of indexers, the phosphorylated α indexer set, the nonphosphorylated γ indexer set, the phosphorylated Bam indexer set and the nonphosphorylated directional BamCC indexer mix set, were employed during the development of the DNA indexing model system. The features of the Bam and BamCC indexer sets are detailed in **Section 2.3.9**.

Two sets of indexers, α and γ, were used in the first stages of pUC19 model system development (FIGURE 2.3). The 24-nt α indexing strand (FIGURE 2.3A) was 5'-phosphorylated following synthesis, and bore a sequence of the general format 5'-P-NNNNTTGTCTTCGCATCCTGTACC-3'. Any particular α indexing strand oligo thus bore a phosphorylated informative 5' 4-nt cohesive end corresponding to one of the 256 tetranucleotide sequences, and the 20-nt α-primer-binding sequence (5'-TTGTCTTCGCATCCTGTACC-3') common to all indexing strands of the α

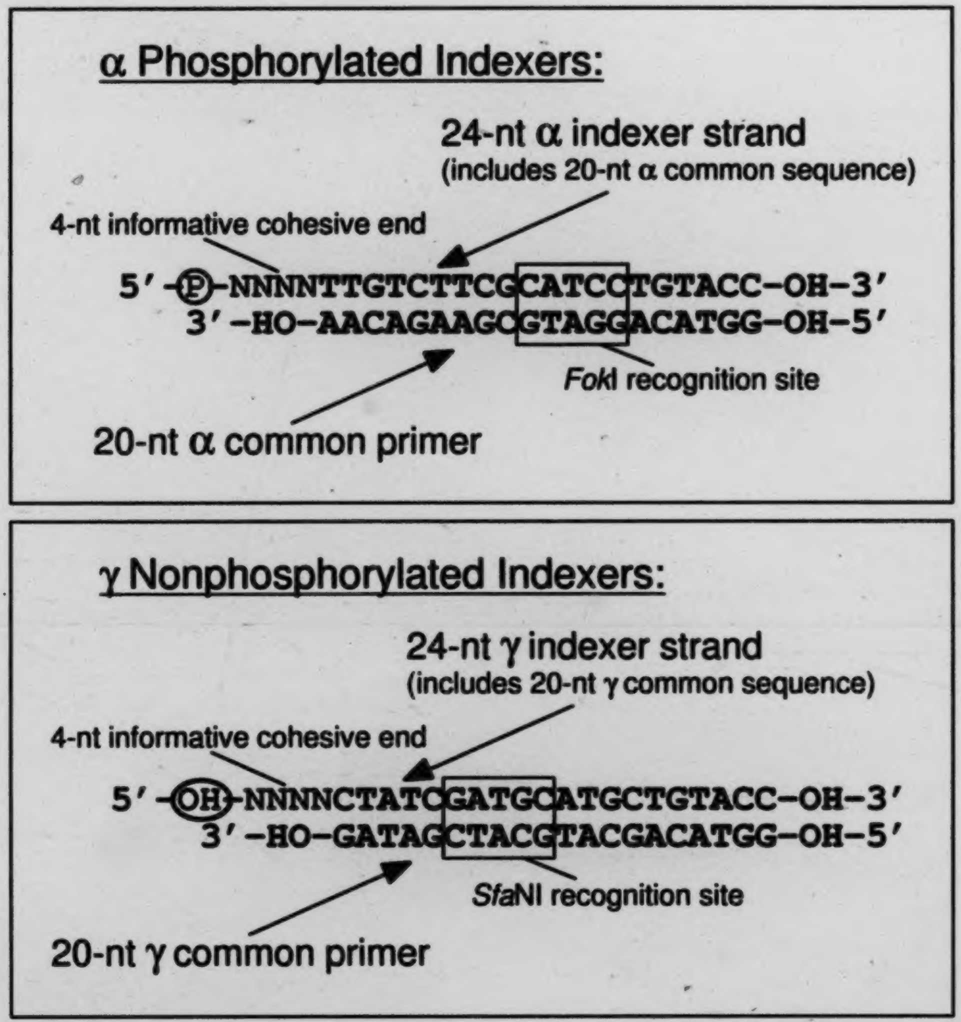

α Phosphorylated Indexers:
24-nt α indexer strand
(includes 20-nt α common sequence)
4-nt informative cohesive end
5'-Ⓟ-NNNNTTGTCTTCGCATCCTGTACC-OH-3'
3'-HO-AACAGAAGCGTAGGACATGG-OH-5'
FokI recognition site
20-nt α common primer
γ Nonphosphorylated Indexers:
24-nt γ indexer strand
(includes 20-nt γ common sequence)
4-nt informative cohesive end
5'-OH-NNNNCTATCGATGCATGCTGTACC-OH-3'
3'-HO-GATAGCTACGTACGACATGG-OH-5'
SfaNI recognition site
20-nt γ common primer

FIGURE 2.3: Features of the α phosphorylated indexer set and the γ nonphosphorylated indexer set.

indexer set. The α primer strand (5'-OH-GGTACAGGATGCGAAGACAA-3') was annealed to each α indexing strand to generate a set of double-stranded phosphorylated indexers. This shorter strand provides the 3'-hydroxyl group required by DNA ligase to form a covalent bond between the primer strand of the indexer and the 5'-phosphate on the 4-nt cohesive end of the target fragment. (A second subset of α indexers with nonphosphorylated indexing strands was also constructed for use in experiments involving the single-primer P/NoP indexing strategy, as described in **Section 2.3.5**) A feature of the double-stranded core sequence of the α indexers is the presence of the recognition sequence 5'-GGATG-3' for the Type IIS restriction endonuclease *Fok*I starting at base position 7 from the 5' end of the α primer strand. The orientation and positioning of this recognition sequence was designed to permit the removal of α indexers from target fragments to which they had been ligated (or from amplified indexing templates derived from these target fragments) by digestion with *Fok*I.

The 24-nt γ indexing strand (FIGURE 2.3B) remained unphosphorylated following synthesis, and bore a sequence of the general format 5'-OH-NNNNCTATCGATGCATGCTGTACC-3'. Any particular γ indexing strand oligo thus bore a nonphosphorylated informative 5' 4-nt cohesive end corresponding to one of the 256 tetranucleotide sequences, and the 20-nt γ-primer-binding sequence (5'-CTATCGATGCATGCTGTACC-3') common to all indexing strands of the γ indexer set. The γ primer strand (5'-GGTACAGCATGCATCGATAG-3') was annealed to each γ indexing strand to generate a set of double-stranded nonphosphorylated indexers. A feature of the double-stranded core sequence of the γ indexers is the presence of the recognition sequence 5'-GCATC-3' for the Type IIS restriction endonuclease *Sfa*NI starting at base position 11 from the 5' end of the α primer strand. The orientation and positioning of this recognition sequence was designed to permit the removal of γ indexers from target fragments to which they had been ligated (or from amplified indexing templates derived from these target fragments) by digestion with *Sfa*NI.

2.3.2 Digestion of pUC19 DNA by *Fok*I restriction endonuclease

While complete target DNA digestion is vital to meaningful DNA indexing analysis, the selected reaction conditions must not sacrifice digest quality for digest completion. Degradation of either double-stranded DNA or of single-stranded informative cohesive ends, and nonspecific cleavage of target DNA (or "star" activity) by the restriction endonuclease jeopardize digest quality and thus the reliability of indexing data. Use of *Fok*I under overdigestion conditions (greater than five units of enzyme per microgram of DNA) or for prolonged incubations may result in either of these digest quality failure modes. Establishing the maximum amount of *Fok*I restriction endonuclease capable of completely digesting pUC19 DNA without significantly degrading the DNA or resulting in "star" activity was important to the development of the DNA indexing model system. A simple digestion assay was therefore performed to establish the optimal conditions for *Fok*I digestion of pUC19 DNA in preparation for indexing.

Commercially-available *Fok*I restriction endonuclease is produced by overexpression of the enzyme from an *E. coli* strain carrying the cloned *fokI* gene from *Flavobacterium okeanokoites* [111]. The supplier lists the recommended incubation temperature for digestion of DNA by *Fok*I as 37°C. Although *E. coli* exhibits optimal growth at 37°C, the physiological temperature of the species which it has evolved to colonize effectively, the protein encoded by the *fokI* gene has evolved to exhibit maximum stability and performance under substantially different environmental temperatures. *F. okeanokoites*, recently reclassified as *Planococcus okeanokoites*, [112] is a bacterium found in marine, freshwater and soil environments, and is closely related to low-%GC species colonizing quite psychrophilic niches such as hypersaline ponds in the McMurdo Ice Shelf of Antarctica [113, 114]. Laboratory cultures of *P. okeanokoites* typically exhibit optimal growth at 20-25°C [113]. Given the cool temperatures which this organism has evolved to exploit, it is likely that enzymes of its restriction-modification system will demonstrate optimal efficiency under similar conditions. The pUC19 DNA *Fok*I digestion assay was therefore designed to determine if a difference in digestion efficiency was evident between the supplier's recommended incubation temperature of 37°C and a lower incubation temperature of 20°C derived from the typical culture conditions of *P. okeanokoites*.

One unit of *Fok*I restriction endonuclease is defined by the supplier as the amount of enzyme required to completely digest 1 µg of λ DNA in a total reaction volume of 50 µl in one hour in NEBuffer 4 at 37°C (see **Section 2.2.2** for buffer description). While this recommendation regarding the amount of enzyme to be employed per microgram of target DNA for complete digestion was a useful guide in designing the pUC19 digestion assay, evaluation of *Fok*I's digestion efficiency in the context of the pUC19 indexing system was necessary due to differences between the reaction conditions of the supplier's defined system and those of the indexing digestion assay. These differences included incubation temperature, reaction volume, and the number of cutsites per microgram of target DNA. (Bacteriophage λ has 150 *Fok*I sites in roughly 50 kb of DNA sequence. In 1 µg of λ DNA there are approximately 2.9×10^{12} *Fok*I sites, while 1 µg of pUC19 DNA has about 1.7×10^{12} sites. Therefore complete digestion of one microgram of pUC19 requires roughly half as many *Fok*I cleavage events than does the complete digestion of the standard lambda system used by the supplier to define one restriction endonuclease unit.) In each tube of the indexing digestion assay, 1 µg of pUC19 DNA was digested with the appropriate number of units (1, 2, 3, 4, or 8 U) of *Fok*I restriction endonuclease at the appropriate incubation temperature (either 37°C or 20°C) for one hour in a total reaction volume of 20 µl of NEBuffer 4. Digestion was halted by denaturation of the restriction endonuclease by heating the reactions to 65°C for 20 min. Ten microlitres of each digest was analyzed by agarose gel electrophoresis.

The data presented in FIGURE 2.4 demonstrate that digestion of 1 µg of pUC19 DNA with more than 1 U of *Fok*I endonuclease resulted in overdigestion and degradation of the target DNA. The EtBr-stained DNA bands corresponding to the five pUC19 *Fok*I fragments in the 1U/37°C and 1U/20°C digests are intense and defined in comparison to those produced by even a 2-fold excess of enzyme. The

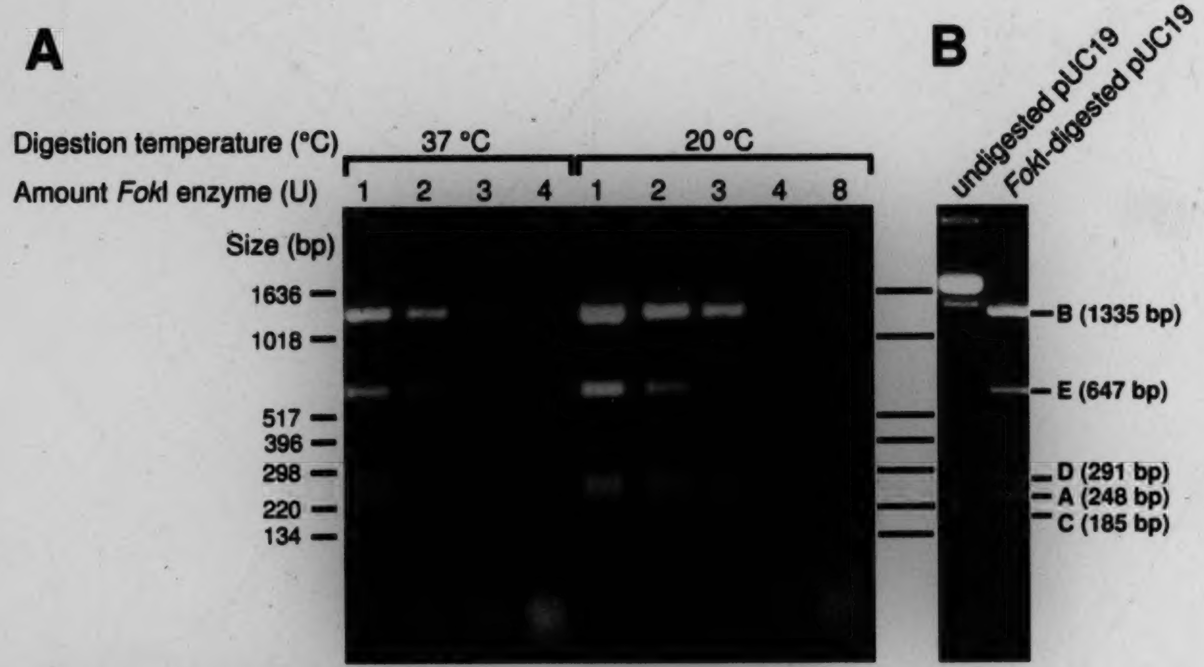

A
Digestion temperature (°C)
37 °C
20 °C
Amount FokI enzyme (U)
1 2 3 4 1 2 3 4 8
Size (bp)
1636
1018
517
396
298
220
134
B
undigested pUC19
FokI-digested pUC19
B (1335 bp)
E (647 bp)
D (291 bp)
A (248 bp)
C (185 bp)

FIGURE 2.4: Digestion of pUC19 by *FokI* restriction endonuclease.
Each digest contained 1 µg of pUC19 DNA and the indicated amount of *FokI* restriction enzyme in a 20-µl reaction. Digestion was halted by heat denaturation of the endonuclease at 65°C for 15 min.
A) Digestion of 1 µg of pUC19 DNA with 1 U *FokI* enzyme at 20°C for 60 min provides complete digestion of the target DNA without substantial DNA degradation or nonspecific "star" activity.
B) Cleavage of pUC19 by *FokI* using the optimized digestion conditions is demonstrated.

presence of nonspecific low-MW DNA degradation products was increased in the 2U/37°C and 2U/20°C digests, evident as smearing towards the lower portion of the lanes, and the intensity of the specifically-cleaved bands was reduced compared to those generated in digests containing a single unit of enzyme. Digestion with 3 U of enzyme produced marked degradation of the target DNA. The only restriction fragment present in undegraded form in sufficient quantity to produce a defined band was the largest, and thus the most-intensely stained, pUC19 *Fok*I fragment. A fourfold excess of enzyme resulted in complete nonspecific degradation of the target DNA.

Further analysis of the data indicates that the quality of DNA restriction fragment populations generated by *Fok*I digestion was greater for digests incubated at 20°C than for *Fok*I digests incubated at 37°C. Each lane of the gel contained the same amount of DNA. In each digest of the assay, pUC19 was completely digested by *Fok*I, as was evident from the absence of undigested pUC19 plasmid and (more conclusively) from the absence of any clearly-defined unidentified bands resulting from incomplete cleavage of a particular pUC19 *Fok*I site. If complete digestion had been achieved without nonspecific cleavage of pUC19, the intensity of the band sets in each lane would have been identical in intensity. The intensity of the bands in the gel lanes containing DNA digested at 20°C was greater than that of the equivalent bands in digests containing the same amount of enzyme but incubated at 37°C. Additionally, the amount of DNA present as nonspecific low-MW DNA degradation products in smears towards the lower portion of lanes was greater for 37°C digests compared to 20°C digests with the same amount of enzyme. The nonspecific DNA products became smaller on average with increasing amounts of enzyme, reflecting higher levels of degradative activity: however, a particular amount of enzyme generated smaller degradation products (and thus indicated a higher level of degradative activity) when incubated at 37°C than the same amount of enzyme incubated at 20°C. For example, 4 U of *Fok*I almost completely degraded 1 μg of pUC19 in one hour at 37°C, while to achieve a similar level of degradation in the same time at 20°C required more than twice that amount of enzyme. Clearly, the degradative and non-specific cleavage activities of *Fok*I endonuclease are more pronounced during 37°C incubations than at 20°C. One hypothetical model for 0this loss of specificity might be decreased stability of the DNA-recognition domain of the enzyme, or of the "piggyback" conformation that normally sequesters the cleavage domain of *Fok*I from the DNA substrate [76, 82]. Overall, the results of this indexing digestion assay demonstrate that incubation of 1 μg of pUC19 DNA with 1 U *Fok*I endonuclease at 20°C for 60 min provides complete digestion of the target DNA without substantial DNA degradation or nonspecific "star" activity.

2.3.3 Amplification of an indexed fragment following target-specific ligation of indexers

The capability of DNA indexing to selectively amplify a specific target fragment from a DNA digest following ligation of cohesive-end-specific indexers to the target fragment was demonstrated in a simple proof-of-principle experiment. The pUC19 *Fok*I Fragment D was selected as an indexing target. Fragment D is 291 bp in length, including its two four-base cohesive ends exhibiting the sequences CTTT and TAAG. Neither of these end sequences are present on any other *Fok*I fragment of pUC19. Three ligation reactions were assembled in 1x T4 DNA ligase buffer at a total volume of 20 μl, each containing 300 ng of *Fok*I-digested pUC19 (containing equal numbers of each of the five pUC19 *Fok*I fragments) and 40 U T4 DNA ligase. Ligation 1 contained 50 fmol of the phosphorylated indexer P-AAAGxα (complementary to the CTTT cohesive end sequence of the target fragment), Ligation 2 contained 50 fmol of the phosphorylated indexer P-CTTAxα (complementary to the TAAG cohesive end sequence of the target fragment), and Ligation 3 contained 50 fmol of each indexer. The ligations were incubated at 16°C for 60 min, and then terminated by denaturing the ligase at 65°C for 20 min. A 2-μl aliquot from each ligation reaction was amplified for 30 cycles of PCR using 40 pmol of the α common primer as the only primer. Aliquots of the amplification reactions were analyzed by agarose gel electrophoresis (FIGURE 2.5). Neither the amplification of Ligation 1 nor

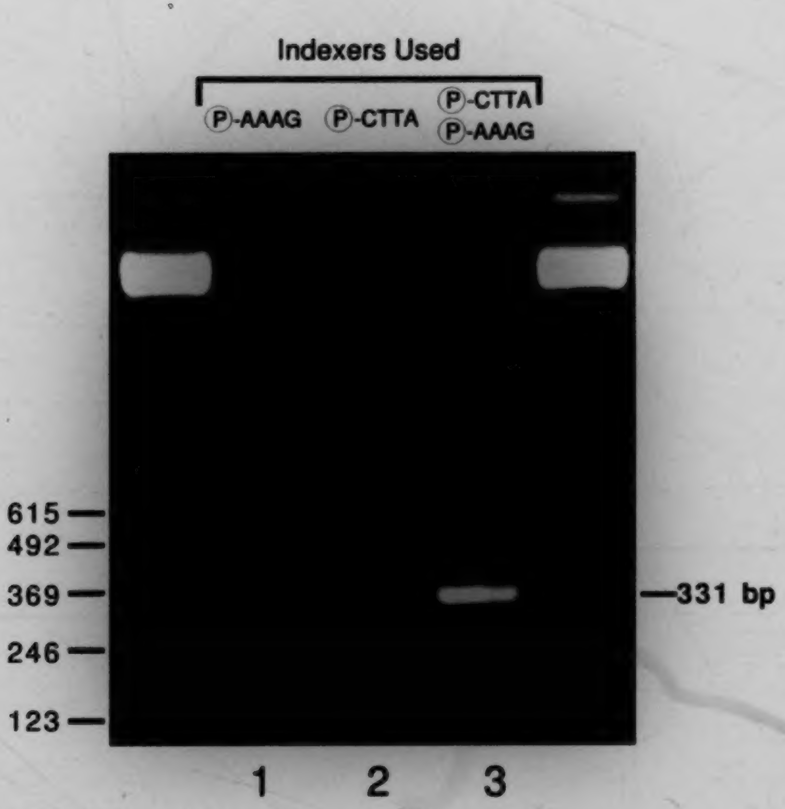


FIGURE 2.5: Target-specific ligation of indexers permits amplification of the indexed fragment.
Ligations containing 300 ng of *Fok*I-digested pUC19 and 40 U T4 DNA ligase were assembled in T4 DNA ligase buffer for a total volume of 20 µl. Ligation 1 contained 50 fmol P-AAAGxα, Ligation 2 contained 50 fmol P-CTTAxα and Ligation 3 contained 50 fmol of each indexer. The ligations were incubated at 16°C for 60 min and terminated by denaturing the ligase at 65°C for 20 min. A 2-µl aliquot of each ligation was amplified for 30 cycles of PCR using 40 pmol α primer, and the results were analyzed by agarose gel electrophoresis.

Ligation 2 resulted in a product. This showed the α primer to be incapable of generating a PCR product from "raw" *Fok*I-digested pUC19 DNA. It also demonstrated that the ligation of a single complementary indexer to one cohesive end of a target fragment was insufficient to allow amplification of that fragment, as there were no PCR products from the ligations containing target fragment that had been indexed with either the P-AAAGxα or the P-CTTAxα indexer alone. The amplification reaction derived from Ligation 3, however, contained a single amplified band of approximately 330 bp in size. The amplified fragment's size was consistent with the predicted length of 291 bp for Fragment D plus 40 bp contributed by the indexers. In other words, a PCR product was observed only when both indexers complementary to each of the target's cohesive end sequences were present in the ligation.

2.3.4 Single-primer P/P indexing of the pUC19 model system

An indexing ligation of a complex *Fok*I DNA digest containing two specific indexer sequences (Indexers 1 and 2) will produce multiple fragments with indexers at both ends. Three classes of fragments will be legitimately represented among the amplification products of such a ligation: one fragment class tagged by one of each specific type of indexer present in the reaction (i.e. Indexer 1 on one end and Indexer 2 on the other; the "intended" target fragment class), and two fragment classes each tagged by two copies of the same indexer (i.e. a class with Indexer 1 legitimately ligated to both ends of the fragments and a class with Indexer 2 legitimately ligated to both ends of the fragments). Fragments falling into one or other of the latter two classes carry the same sequence on both of their cohesive ends. This type of fragment is referred to as a *repeated-end fragment*.

Repeated-end fragments are more likely to be found in complex DNA digests in which the numbers of fragments generated approaches or exceeds the number of indexing fragment classes theoretically produced by digestion with the particular indexing restriction enzyme employed. (The probability that a particular fragment in a digest of a given complexity will be a repeated-end fragment with a given cohesive end sequence is determined by the probability that any one *Fok*I cleavage site will generate a given cohesive end sequence times the probability that the next *Fok*I cleavage site along the DNA will generate the identical cohesive end sequence multiplied by the number of repeated-end fragment classes generated by the indexing restriction endonuclease employed.) The need to resolve repeated-end fragment classes from their related unique-sequence fragment classes complicates indexing analysis of large DNAs by (potentially) tripling the number of fragments ligated and amplified in each indexing set. *Fok*I digestion of a "typical" mammalian genome (2-4 gigabases) would be expected to produce about 6 x 10^6 fragments, with each of the 32 896 indexing fragment classes generated by this enzyme containing about 200 fragments, on average. Indexing of any one fragment class defined by distinct end sequences (***unique-ended fragments***) by two phosphorylated indexers (the *P/P approach*) would be expected to generate about 600 indexed fragments in each ligation. Indexing-based high-resolution mapping of similar genomes, in preparation for directed sequencing for example, requires a method of discriminating between indexed fragments of the intended target class and the repeated-end fragment classes. This difficulty is only avoided entirely in those cases in which the intended target class to be analyzed is a repeated-end fragment.

Contrary to what would be expected for such a small DNA molecule, two of the five *Fok*I sites in pUC19 generate identical cohesive end sequences. As a result, the pUC19 model system can produce two repeated-end amplicons through multiple ligation events, one double-fragment amplicon created by a re-ligation of Fragment A and Fragment B, and a triple-fragment amplicon produced by re-ligation of Fragments C, D and E. This characteristic makes the pUC19 model system attractive for the development of indexing strategies which are capable of resolving artifactual fragment amplification due to repeated-end fragments.

The *single-primer P/P indexing* strategy is the most basic form of DNA indexing. Using this approach, two phosphorylated indexers bearing a single common primer-binding sequence are ligated to complementary end sequences on one of each of the five pUC19 *Fok*I restriction fragments. FIGURE 2.6A is a schematic

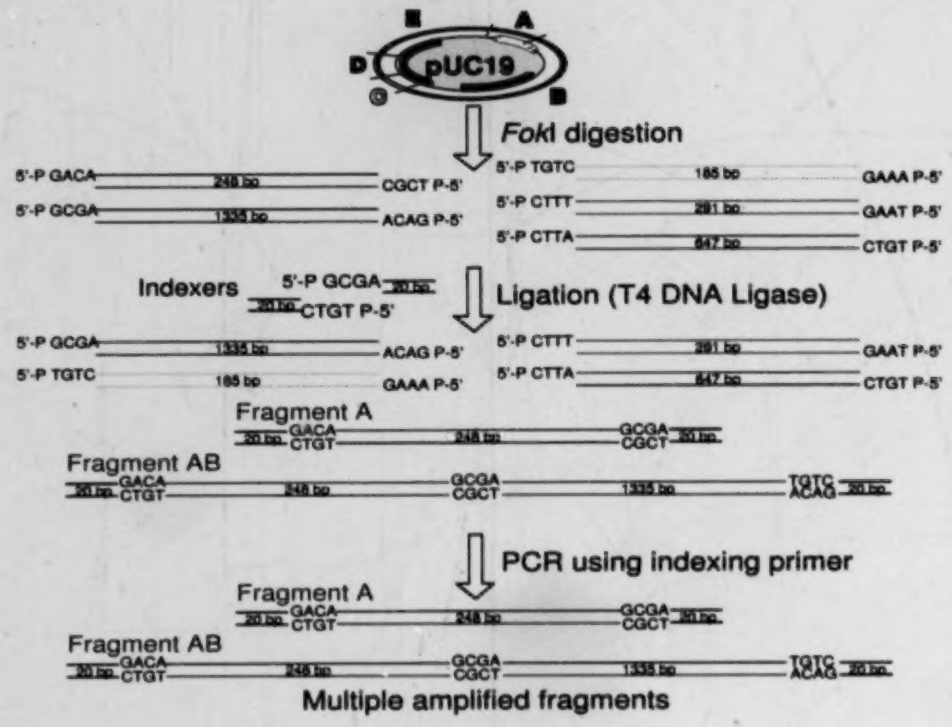

pUC19
A
B
D
Fokl digestion
5'-P GACA
248 bp
CGCT P-5'
5'-P GCGA
1335 bp
ACAG P-5'
5'-P TGTC
165 bp
GAAA P-5'
5'-P CTTT
291 bp
GAAT P-5'
5'-P CTTA
647 bp
CTGT P-5'
Indexers
5'-P GCGA
20 bp
CTGT P-5'
Ligation (T4 DNA Ligase)
Fragment A
GACA
CTGT
GCGA
CGCT
Fragment AB
TGTC
ACAG
PCR using indexing primer
Multiple amplified fragments

FIGURE 2.6: Single-primer P/P indexing of pUC19 by α indexers.
A) Schematic flow diagram outlining single-primer P/P indexing of Fragment A. See text for details.

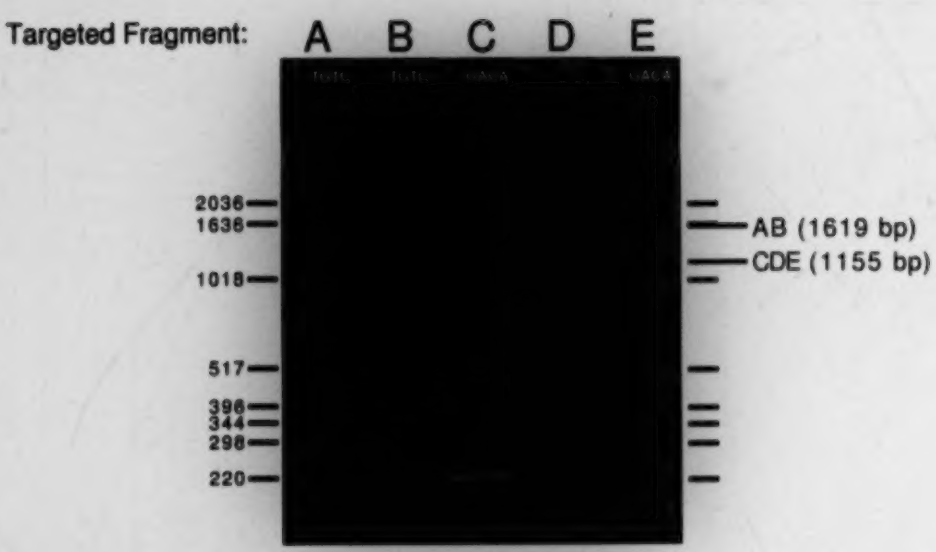

Targeted Fragment:
A
B
C
D
E
2036
1636
1018
517
396
344
298
220
AB (1619 bp)
CDE (1155 bp)

FIGURE 2.6: Single-primer P/P indexing of pUC19 by α indexers.
B) Single-primer P/P indexing of Fragments A to E of pUC19. Ligations targeting one of the five pUC19 *Fok*I fragments each contained two α P-indexers (50 fmol/indexer), 300 ng of *Fok*I-digested pUC19 and 40 U T4 DNA ligase. Reactions were incubated at 16°C for 1 h, followed by heat denaturation of the ligase. PCR and agarose gel electrophoresis were performed as previously described.

representation of the mechanism underlying the amplification of the AB double fragment amplicon as a byproduct of the amplification of Fragment A using the single-primer P/P indexing strategy. Indexers P-GCGAxα and P-TGTCxα are added to a ligation reaction containing FokI-digested pUC19 DNA and T4 DNA ligase. Accurate joining of indexers to their complementary cohesive end sequences permit not only the ligation of P-GCGAxα to the TCGC cohesive end and P-TGTCxα to the GACA cohesive end of Fragment A, but also the ligation of P-TGTCxα to the GACA cohesive end of Fragment B. An additional legitimate ligation event, religation of the TCGC end of Fragment A with Fragment B's GCGA end sequence, assembles a DNA molecule with P-TGTCxα on both ends of the religated AB fragment. This repeated-end double-fragment indexing amplicon is legitimately amplified by an indexing PCR reaction containing the α common primer, despite being of a different fragment class than the desired indexing target Fragment A. The same fragment will be amplified by single-primer P/P indexing of Fragment B. (Assembly of the repeated-end CDE triple fragment proceeds by a similar mechanism in ligations containing the P-GACAxα indexer.)

The pUC19 model system was indexed using the single-primer P/P indexing strategy in order to demonstrate the targeted amplification of each of the pUC19 *Fok*I fragments, and to observe the generation of non-targeted repeated-end fragments (FIGURE 2.6B). To ensure that all five target fragments would be indexed effectively, and to provide a benchmark against which efforts to optimize indexing reaction conditions might be evaluated, ligation was performed under nonstringent conditions expected to provide enhanced ligation efficiency. Five indexing ligation reactions were assembled, each containing two phosphorylated α indexers (50 fmol/ligation of each indexer) targeting one of the five pUC19 *Fok*I fragments, and 300 ng of *Fok*I-digested pUC19.

As expected, all five target fragments were correctly amplified by the indexing reactions which targeted them. Amplification of the repeated-end AB double-fragment from indexing reactions containing the P-TGTCxα indexer (lanes A and B) was observed as anticipated, as was the amplification of the CDE triple-fragment from reactions containing the P-GACAxα indexer (lanes C and E). A nontargeted misligation product corresponding to Fragment E was evident in the indexing reaction targeting Fragment C. In this ligation, the P-GACAxα indexer correctly targeted the TGTC end sequence present on both the C and E fragments (and, as a consequence, on both ends of the CDE religation molecule). However, the other indexer present in the reaction (P-CTTTxα) should target only the C fragment's AAAG end under non-permissive conditions. For Fragment E to be ligated to an indexer on both ends, its CTTA end must be misligated to either P-GACAxα or P-CTTTxα, or a more radical form of misligation may be taking place due to the forcing DNA concentrations and permissive incubation temperatures used in the indexing reactions.

These results demonstrate the successful targeted amplification of pUC19 Fragments A to E using the single-primer P/P strategy, while drawing attention to the challenge posed to indexing systems by repeated-end fragments. In addition, the amplification of misligated fragments due to the permissive ligation conditions used in this experiment demonstrates the need to establish indexing reaction conditions that permit efficient joining of indexers to target fragment ends while ensuring high ligation fidelity. The pUC19 model system was subsequently applied to the development of indexing protocols providing resolution of the challenges identified by this experiment.

2.3.5 The P/NoP indexing strategy

As discussed above, ligation of a complex *Fok*I DNA digest with two specific indexer sequences will target three classes of fragments: one unique-ended fragment class, and two collateral classes of repeated-end fragments. The effect of this tripling in complexity of the indexing analysis of large fragment sets may be reduced through the use of nonphosphorylated indexers. The *P/NoP indexing* strategy (FIGURE 2.7) employs both phosphorylated indexers (*P-indexers*) and nonphosphorylated indexers (*NoP indexers*) in a manner that reduces or eliminates the amplification of repeated-end fragments.

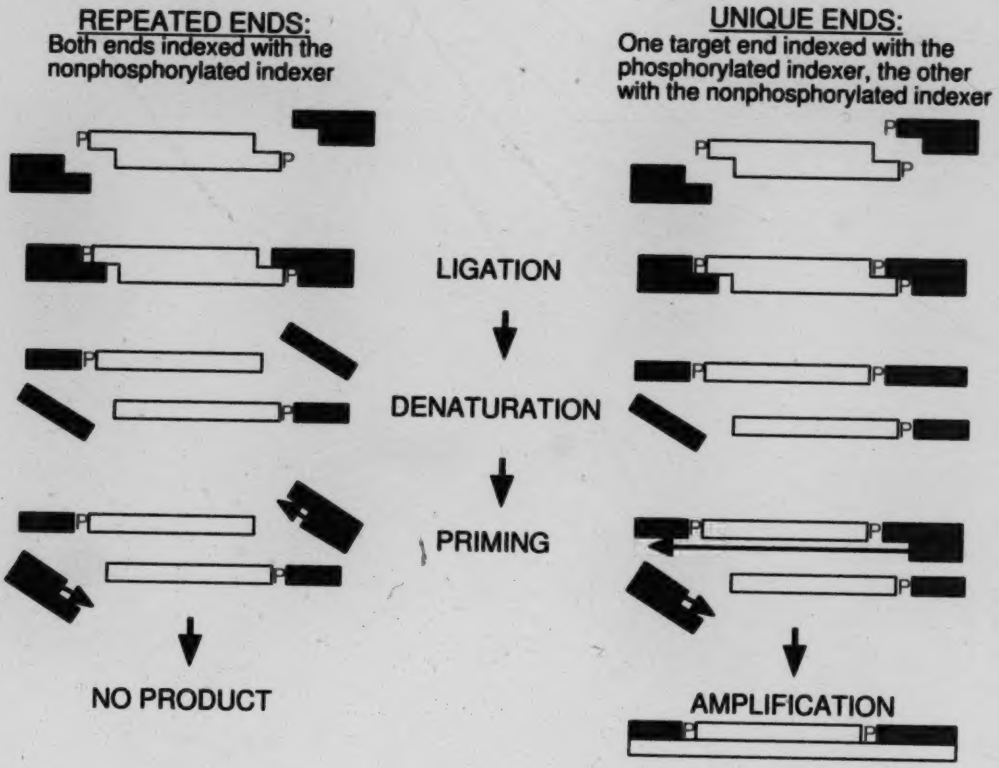

REPEATED ENDS:
Both ends indexed with the nonphosphorylated indexer
UNIQUE ENDS:
One target end indexed with the phosphorylated indexer, the other with the nonphosphorylated indexer
P
LIGATION
DENATURATION
PRIMING
NO PRODUCT
AMPLIFICATION

FIGURE 2.7: P/NoP indexing circumvents the collateral amplification of repeated-end fragment classes.
See text for description.

In P/NoP indexing, a P-indexer and a NoP indexer, each complementary to one of the two sequences of the target fragment class, are ligated to a complex DNA digest. Three ligation events are possible between the double-stranded indexers and a unique-ended fragment of the target class following accurate base-pairing of indexers to the target: the 5'-phosphate groups present on the digested ends of the restriction fragments are joined to the 3'-hydroxyl groups of the primer strands of both the P-indexer and the NoP indexer, and the 5'-phosphate group of the P-indexer's indexing strand and one of the recessed 3'-ends of the target. However, the indexing strand of the NoP indexer is not ligated to the target due to the lack of a 5'-phosphate group. During the denaturation step of PCR, the unligated NoP indexing strand dissociates from the primer strand (which is ligated to the target). As the indexing strand bears the core sequence to which the indexing primer must bind to initiate strand synthesis, synthesis is not primed from that end of the indexed molecule. DNA synthesis is primed from the indexing primer binding to the core sequence of the P-indexer at the opposite end of the target. Amplification of the P/NoP indexing target is therefore linear in the first cycle of PCR. In subsequent cycles, amplification proceeds exponentially, as the primer-binding sequence at the NoP-indexed end of the target is reconstituted using the indexer primer sequence as a template.

In contrast, only two ligation events are permitted between a repeated-end fragment and two NoP indexers targeting its cohesive end sequence. The 5'-phosphate groups of the restriction fragment are joined to the 3'-hydroxyl groups of the primer strands of both incoming NoP indexers. However, neither of the non-phosphorylated indexing strands are ligated to the target. As a result, both primer-binding sequences are lost upon denaturation in PCR. Without a primer-binding sequence at either end of the molecule, no DNA synthesis is possible, and amplification of the repeated-end fragment will not occur.

In uncharacterized complex DNA digests, use of the P/NoP system reduces the number of collateral amplification products (repeated-end fragments) twofold by eliminating the amplification of the class of repeated-end fragments targeted by the sequence of the NoP indexer used in ligation. Amplification of the unique-ended target fragment class and the repeated-end fragment class targeted by the P-indexer sequence is enabled. Identification of products as members of the unique-ended target class is possible by reversing the assignation of cohesive end sequence between P-indexer and NoP indexer ("switching the polarity") and repeating the experiment. This will permit P/P amplification of the repeated-end class previously prohibited by NoP/NoP indexers, and prohibit amplification of the previously-observed repeated-end fragment class. Fragments amplified in both experimental sets are members of the unique-ended target fragment class. In instances in which well-characterized DNA mixtures are being indexed (e.g. in preparation for probe generation), it is often possible to select the P-indexer and NoP indexer sequences appropriately to eliminate repeated-end amplification entirely.

2.3.6 Establishing ligation conditions for DNA indexing

Reaction conditions for the joining of indexers to target *Fok*I fragments by T4 DNA ligase were established to provide an appropriate balance between ligation fidelity and ligation efficiency. Indexers were ligated to *Fok*I-digested pUC19 target fragments at varying temperatures, at different concentrations of T4 DNA ligase, and using different amounts of digested DNA (FIGURE 2.8). Ligation reaction products were amplified by PCR and analyzed by agarose gel electrophoresis. Ligation fidelity was evaluated by the absence of amplification products corresponding to untargeted digest fragments. Ligation efficiency was evaluated by the amplification of accurately-indexed target fragments.

Each indexing ligation contained the relevant amount of *Fok*I-digested pUC19 DNA, the appropriate amount of T4 DNA ligase, and 50 fmol of a phosphorylated indexer and 50 fmol of a nonphosphorylated indexer in 20 µl of ligase reaction buffer. Each indexer was selected to target one of the two cohesive end sequences of a particular *Fok*I fragment of pUC19. In order to eliminate the amplification of repeated-end fragments, a single-primer P/NoP strategy was employed. If the target fragment possessed an end sequence which was also present on another pUC19 fragment in the digest, the nonphosphorylated indexer was used to target that cohesive end sequence.

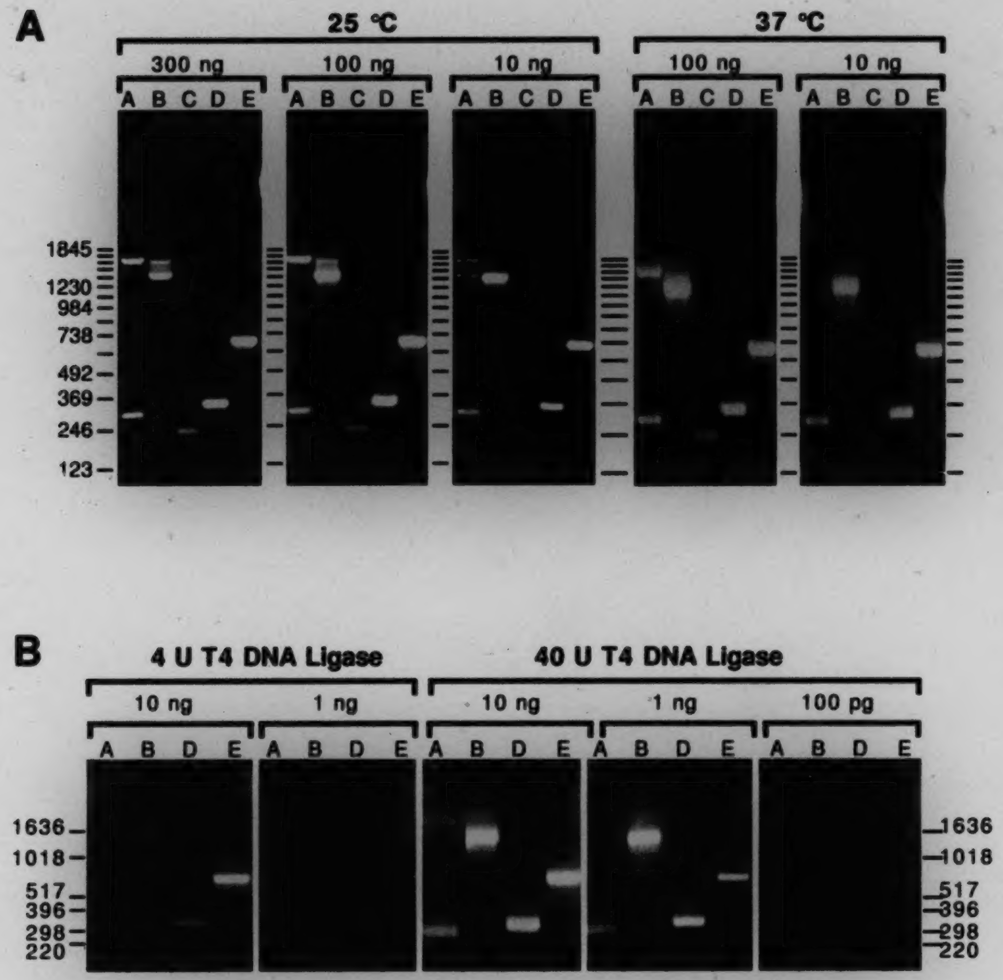

A
25 °C
37 °C
300 ng
100 ng
10 ng
100 ng
10 ng
A B C D E
1845
1230
984
738
492
369
246
123
B
4 U T4 DNA Ligase
40 U T4 DNA Ligase
10 ng
1 ng
10 ng
1 ng
100 pg
A B D E
1636
1018
517
396
298
220

FIGURE 2.8: Determination of optimal ligation conditions for DNA indexing. Indexing reactions targeting Fragment A contained indexers **P**-GCGAxα and **OH**-TGTCxα; for Fragment B, **P**-TCGCxαand **OH**-TGTCxα; for Fragment C, **P**-CTTTxα and **OH**-GACAxα; for Fragment D, **P**-AAAGxα and **OH**-CTTAxα; for Fragment E, **P**-TAAGxα and **OH**-GACAxα.
A) Each indexing ligation contained the relevant amount of *Fok*I-digested pUC19 DNA, 40 U T4 DNA ligase, and 50 fmol each of a P-indexer and a NoP indexer in a 20-µl reaction. Ligations were incubated for 1 h at the indicated temperature, and the ligase denatured at 65°C for 15 min.
B) Each ligation contained the relevant amount of *Fok*I-digested pUC19 DNA, the noted amount of T4 DNA ligase, and 50 fmol each of a P-indexer and a NoP indexer in a 20-µl reaction. All ligations in B were incubated for 1 h at 37°C, and the ligase denatured at 65° C for 15 min. Two microlitres of each ligation was added to a 50-µl PCR reaction containing native *Taq* DNA polymerase and utilizing 40 pmol of α primer. PCR and agarose gel electrophoresis were performed as previously described.

The unique end of the target fragment in each case was ligated to a phosphorylated indexer. For indexing reactions targeting Fragment A, the indexers selected were P-GCGAxα and OH-TGTCxα. For indexing reactions targeting Fragment B, the indexers selected were P-TCGCxα and OH-TGTCxα. For indexing reactions targeting Fragment C, the indexers selected were P-CTTTxα and OH-GACAxα. For indexing reactions targeting Fragment D, the indexers selected were P-AAAGxα and OH-CTTAxα. For indexing reactions targeting Fragment E, the indexers selected were P-TAAGxα and OH-GACAxα. All ligations were incubated for 1 h at the appropriate temperature, and the reaction was halted by denaturation of the ligase at 65°C for 15 min. Two microlitres of each completed ligation reaction was added to a 50-µl PCR reaction containing native *Taq* DNA polymerase and utilizing 40 pmol of α indexing primer as the sole common primer. Amplification proceeded for thirty cycles under standard indexing PCR conditions.

In the first part of this experiment (FIGURE 2.8A), the ligation fidelity and efficiency provided at several incubation temperatures was evaluated for a range of DNA digest concentrations with a fixed concentration (40 U) of T4 DNA ligase. Incubation temperatures of 16°C or below were found to be permissive for the joining of inaccurately base paired cohesive end sequences of indexers and digest fragments (or "misligation") (data not shown). One set of ligations was performed at 25°C using large amounts (300 ng) of *Fok*I-digested pUC19 to ensure indexing and amplification of all targeted fragments. At 25°C, misligated fragments were observed for some indexer sequences at all DNA concentrations tested, as evidenced by the presence of untargeted PCR products in certain indexing reactions (e.g. the presence of high-molecular-weight amplification products in PCRs derived from Fragment A ligations containing 300 ng, 100 ng, and 10 ng of *Fok*I-digested pUC19 DNA). Ligations containing high concentrations of digested pUC19 DNA (5 ng/µl, or 100 ng/ligation) generated misligated products in some reactions even at 37°C, a temperature at which base pairing of unligated 4-nt cohesive ends is unstable even for correctly-paired sequences. The less forcing conditions found at lower DNA concentrations reduced the opportunity for misligations to occur. Indexing reactions performed with 10 ng *Fok*I-digested pUC19 (500 pg/µl) at 37°C contained no misligation products, and provided the ligation efficiency required to permit efficient amplification of all pUC19 target fragments.

To determine the appropriate amounts of ligase and *Fok*I-digested pUC19 DNA for target fragment indexing at 37°C, indexing reactions containing 10 ng, 1 ng and 100 pg of DNA were ligated using 4 U or 40 U of enzyme (FIGURE 2.8B). In general, 4 U of T4 DNA ligase did not efficiently ligate indexers to target fragments. Even at DNA concentrations of 500 pg/µl, some target fragments were inefficiently indexed and misligation products were produced. At 50 ng/µl of DNA, some ligations employing 4 U of enzyme failed entirely to generate target products, though for other indexing sequence combinations a low level of product was obtained. For reactions with DNA concentrations below 50 ng/µl, 4 U of ligase was insufficient to produce indexed target fragments for any of the indexer sequences tested (data not shown). For indexing reactions performed using 40 U of ligase, efficient ligation of all indexing ends tested was demonstrated for DNA concentrations of 500 pg/µl and 50 pg/µl. Some target fragments were indexed and amplified from ligations containing 5 pg/µl (100 pg/ligation) of *Fok*I-digested pUC19; however, this could not be demonstrated for all tested indexing end sequences. From these results, the reaction conditions appropriate for use in the development of the pUC19 DNA indexing model system were determined to be 50 pg/µl (1 ng/ligation) *Fok*I-digested pUC19 DNA, 50 fmol of each indexer, and 40 U of T4 DNA ligase in a 20-µl reaction volume, incubated at 37°C for 1 h with subsequent denaturation of the ligase at 65°C for 15 min. The ability of the P/NoP indexing strategy to eliminate the production of false positives due to repeated-end amplification was also successfully demonstrated.

2.3.7 Directionality as a requirement for cycle sequencing of indexed template fragments.

The ability to prime DNA synthesis from one end of a DNA template or other but not both simultaneously *(directionality)* is a prerequisite for cycle sequencing of amplified indexed fragments. The use of a single common primer to initiate DNA synthesis from both ends of an indexed fragment during amplification eliminates directionality in a sequencing reaction. Alternate indexing strategies need to be employed to permit sequencing of indexing fragments. Indexed sequencing templates may be generated by ligating indexers of different sets (i.e. bearing different core sequences) to either end of a target fragment. The use of two priming sequences to amplify a target fragment permits directional initiation of DNA synthesis, and allows sequencing of an indexed DNA template to proceed from one end of the molecule or the other. Consequently, early attempts to provide directionality in indexed template amplification focused on the use of two indexer sets, each with a different core sequence and common primer. The α P-indexer set and the γ NoP indexer set were applied to the *double-primer P/NoP indexing* of pUC19 in preparation for cycle sequencing of the indexed amplicons. The use of two different priming sequences in an indexing reaction produced an artifact characterized as a nontypical form of primer-dimer (PD). This artifact was capable of out-competing the amplification of correctly-indexed target fragments. As PD formation in two-primer indexing PCR reactions reduced or prohibited target fragment amplification, other strategies bestowing directionality upon indexed fragment priming were developed and evaluated.

2.3.8 Generation and description of indexing PD artifact

Primer-dimer in its general form is a non-specific artifact of amplification [115-117] generated through template-independent primer interactions that reduce the efficiency of target product amplification through competition for nucleotides and enzyme [115, 118, 119]. They are not derived from template DNA and can complicate experimental analysis. Archetypical primer-dimer formation is a result of two primers used in a PCR reaction sharing some degree of 3'-end complementarity, annealing and priming off of one another, and generating a very short PCR product with a specific amplification efficiency relative to the reaction's target amplicon(s). Complementarity of just one nucleotide between amplimer 3'-ends can permit the generation of PD artifacts [95]. Primer-dimer formation is favoured by such conditions as low annealing temperatures, cold starts [116, 120], high enzyme and primer concentrations [121], multiple primers [122], and primer 3'-end complementarity. Proper selection of primer sequences, stringent amplification temperatures, or the use of polymerases inactivated with anti-*Taq* antibodies [117, 123, 124] have been demonstrated to reduce the occurrence of this general form of primer-dimer.

An unusual form of primer-dimer was observed in indexing amplification reactions involving two primers of unrelated sequence (FIGURE 2.9). The products of five indexing reactions corresponding to the double-primer P/NoP indexing of pUC19 by fragments of the α P-indexer set and the γ NoP indexer set were ligated and amplified following protocols previously demonstrated to efficiently amplify the pUC19 target fragments when a single set of indexers were used. Fragment A was indexed with P-GCGAxα and OH-TGTCxγ; Fragment B by P-TCGCxα and OH-TGTCxγ; Fragment C by P-CTTTxα and OH-GACAxγ; Fragment D by P-AAAGxα and OH-CTTAxγ; and Fragment E by P-TAAGxα and OH-GACAxγ. All ligations contained 1 ng of *Fok*I-digested pUC19 DNA and 40 U of T4 DNA ligase, and were incubated for 1 h at 37°C. The reaction was halted by denaturation of the ligase at 65°C for 15 min. Two microlitres of each completed ligation reaction was added to a 50-µl PCR reaction containing antibody-sequestered *Taq* DNA polymerase and utilizing 20 pmol each of the α and γ indexing primers. Following a hot start, amplification proceeded for thirty cycles under standard indexing PCR conditions. The specific amplification efficiency of the 40-bp PD artifact was so highly competitive that the presence of PD in a PCR prevented the amplification of target fragments to levels sufficient for visualization on agarose gel (FIGURE 2.9A).

The PCR products generated from indexing reactions using two unrelated primers were compared with the PCR products produced using a single primer. Several sets of indexers with unrelated primer sequences were evaluated against each other in all combinations of sequence pairings, and in single-primer indexing approaches. Despite the incorporation of preventative measures such as hot starts, elevated annealing temperatures and antibody-sequestered *Taq* DNA polymerase, the

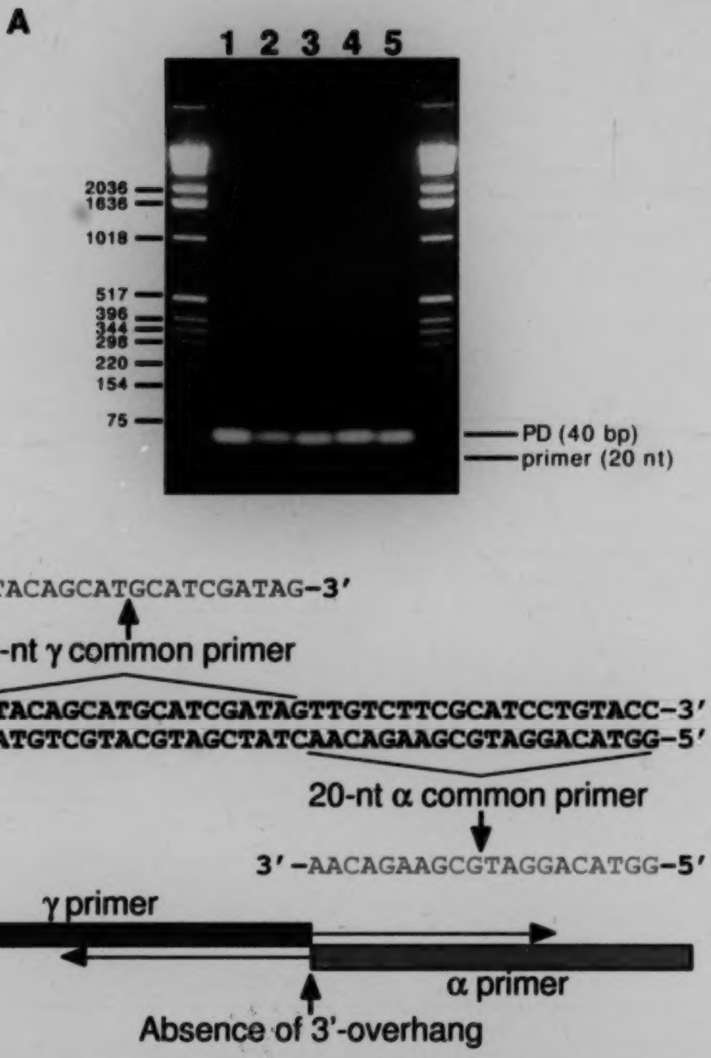

A
1 2 3 4 5
2036
1636
1018
517
396
344
298
220
154
75
PD (40 bp)
primer (20 nt)
B
5'-GGTACAGCATGCATCGATAG-3'
20-nt γ common primer
5'-GGTACAGCATGCATCGATAGTTGTCTTCGCATCCTGTACC-3'
3'-CCATGTCGTACGTAGCTATCAACAGAAGCGTAGGACATGG-5'
20-nt α common primer
3'-AACAGAAGCGTAGGACATGG-5'
γ primer
α primer
Absence of 3'-overhang

FIGURE 2.9: Production and structure of indexing PD artifact.

A) Fragment A was indexed with P-GCGAxα and OH-TGTCxγ; Fragment B by P-TCGCxα and OH-TGTCxγ; Fragment C by P-CTTTxα and OH-GACAxγ; Fragment D by P-AAAGxα and OH-CTTAxγ; and Fragment E by P-TAAGxα and OH-GACAxγ. All ligations contained 1 ng of *Fok*I-digested pUC19 DNA and 40 U of T4 DNA ligase, and were incubated for 1 h at 37° C. A 2-μl aliquot of each ligation was added to a 50-μl PCR reaction containing antibody-sequestered *Taq* DNA polymerase and 20 pmol each of the α and γ indexing primers. Following a hot start, was performed for 30 cycles. The 40-bp PD artifact so effectively out-competes amplification of target fragments that its presence reduced target fragment amplification below the level required for visualization of a band on an agarose gel.

B) Sequence analysis showed that the indexing PD artifact was a 40 bp fragment composed of the 20-nt α primer sequence juxtaposed with the 20-nt γ primer sequence on the opposite strand without 3' overlap of the two primer sequence.

inhibitory PD artifact was observed in the case of every tested combination of different primers irrespective of an 3'-end complementarity between primer sequences. No PD was generated in any single-primer PCR reaction (data not shown).

The PD artifact was isolated, cloned and sequenced (as described in Section 2.2.9) in the hope that determining its structure might shed light on the mechanism by which it was generated, and thus suggest measures by which PD formation might be prevented. Sequence analysis showed that the indexing PD artifact was a 40 bp fragment composed of the 20-nt α primer sequence juxtaposed with the 20-nt γ primer sequence on the opposite strand without 3'-end complementarity or, indeed, without any 3' overlap of the two primer sequences (FIGURE 2.9B). The mechanism by which such a construct could be formed by the known activities of *Taq* DNA polymerase is unclear, as is any remedial action which could be taken to prevent its formation. Recently, evidence was presented of the ability of two thermostable DNA polymerases isolated from bacterial species closely related to *Thermus aquaticus* to synthesize up to 200 kb of linear double-stranded DNA *in vitro* in the complete absence of primer and template DNAs [125, 126]. If this capacity for *de novo* creation of genetic information were shared by the related *Taq* DNA polymerase, a potential mechanism underlying the formation of nontypical PD artifacts could be suggested.

The dependency of primer-dimer generation in DNA indexing reactions on the presence of two or more non-identical primers in the PCR reaction has also been noted by Brownie *et al.* [122] in their study of PCR primers for multiplex amplification, where several pairs of PCR primers are required simultaneously. Brownie *et al.* were unable to generate primer-dimers either with individual primers alone or with similar sequence primers even if they had 3' complementarity. The same study showed that in an analogous PCR, where there was no 3'-complementarity, PDs were still produced. In some instances nucleotides were deleted from the 3' end of one or both primers; in others a seemingly random sequence of nucleotides was inserted between the 3' ends of primers; and in others a string of nucleotides derived from one of the primer sequences was inserted between the primers. Despite extensive sequence analysis of

PD constructs with such an array of unusual characteristics, the authors were unable propose a favoured mechanism promoting PD formation.

Other groups employing indexing approaches have reported similar findings in instances in which two indexing primer sequences were present in the same PCR. These groups have evaluated several methods for reduction of primer-dimer, including size-selective adsorption to glass beads or other matrices, elution from silica gel membranes with guanidine hydrochloride and gel excision, but none of the methods evaluated compensates for the reduced efficiency of PCR reactions containing primer-dimer [127, 128]. A more efficient approach is to develop an indexing approach that sidesteps the formation of the PD artifact entirely, while providing the directionality of indexing amplification required to permit cycle sequencing of indexed target fragments. Elimination of primer-dimer from indexed PCR reactions offers the advantage that DNA sequencing templates produced through such an approach may be sequenced directly from amplification reactions without requiring further purification. When more than one indexed target is amplified in a PCR reaction, the higher target amplification efficiencies possible in the absence of PD contamination result in higher target concentrations, facilitating recovery of gel-purified fragments. As PD formation in two-primer indexing PCR reactions reduced or prohibited target fragment amplification, other strategies bestowing directionality upon indexed fragment priming were developed and evaluated.

2.3.9 Features of the Bam phosphorylated indexers and BamCC nonphosphorylated indexers

The Bam P-indexer set and the BamCC NoP indexer set were employed in the final stage of indexing strategy development using the pUC19 model system (FIGURE 2.10). The 24-nt Bam indexing strand (FIGURE 2.10A) was 5'-phosphorylated following synthesis, and bore a sequence of the general format 5'-P-NNNNTTGTCTTCGGATCCTGTACC-3'. Any particular Bam indexing strand oligo thus bore a phosphorylated informative 5' 4-nt cohesive end corresponding to one of the 256 tetranucleotide sequences, and the 20-nt Bam-primer-binding sequence

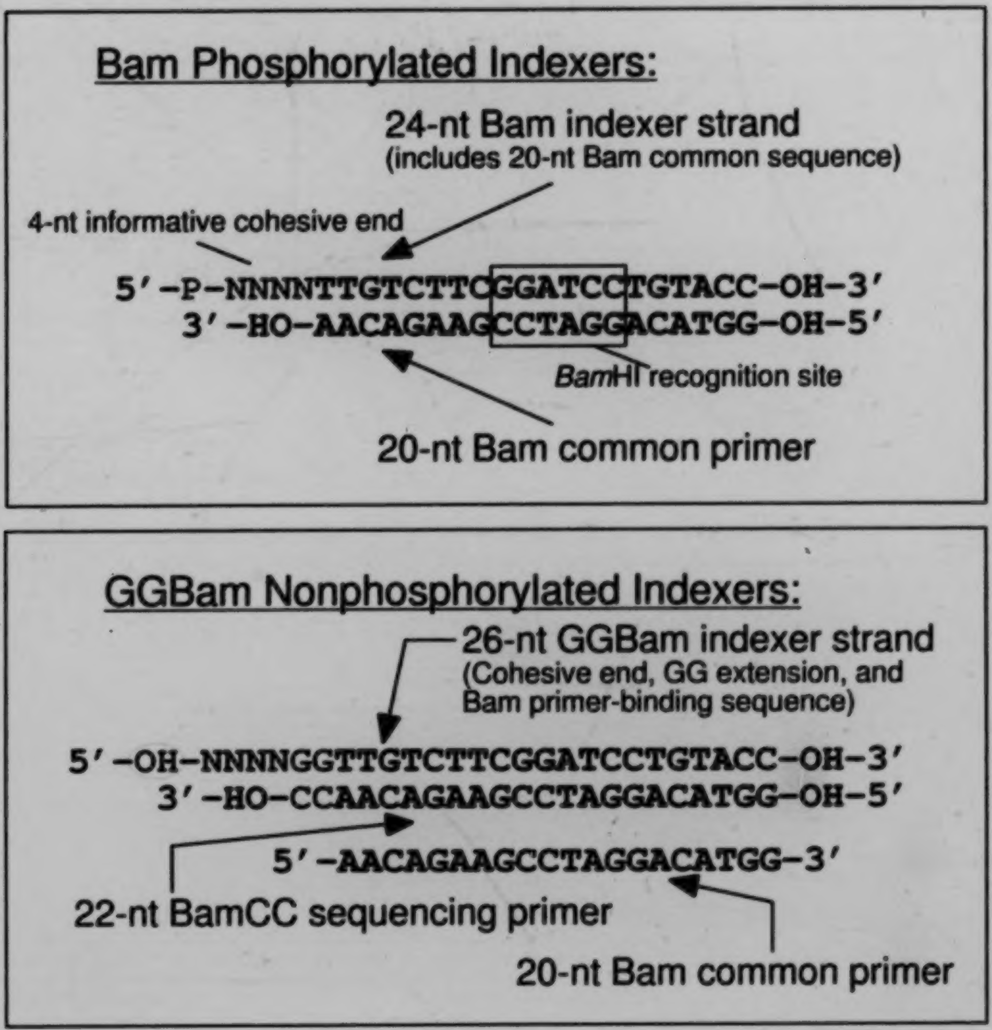

Bam Phosphorylated Indexers:
24-nt Bam indexer strand
(includes 20-nt Bam common sequence)
4-nt informative cohesive end
5'-P-NNNNTTGTCTTCGGATCCTGTACC-OH-3'
3'-HO-AACAGAAGCCTAGGACATGG-OH-5'
BamHI recognition site
20-nt Bam common primer
GGBam Nonphosphorylated Indexers:
26-nt GGBam indexer strand
(Cohesive end, GG extension, and
Bam primer-binding sequence)
5'-OH-NNNNGGTTGTCTTCGGATCCTGTACC-OH-3'
3'-HO-CCAACAGAAGCCTAGGACATGG-OH-5'
5'-AACAGAAGCCTAGGACATGG-3'
22-nt BamCC sequencing primer
20-nt Bam common primer

FIGURE 2.10: Features of the Bam phosphorylated indexer set and the BamCC nonphosphorylated compound indexer set.

(5'-TTGTCTTCGGATCCTGTACC-3') common to all indexing strands of the Bam indexer set. The Bam primer strand (5'-OH-GGTACAGGATCCGAAGACAA-3') was annealed to each Bam indexing strand to generate a set of double-stranded phosphorylated indexers. A feature of the double-stranded core sequence of the Bam indexers is the presence of the recognition sequence 5'-GGATCC-3' for the Type II restriction endonuclease *Bam*HI starting at base position 7 from the 5' end of the Bam primer strand. The orientation and positioning of this recognition sequence was designed to permit the removal of Bam indexers from target fragments to which they had been ligated (or from amplified indexing templates derived from these target fragments) by digestion with *Bam*HI.

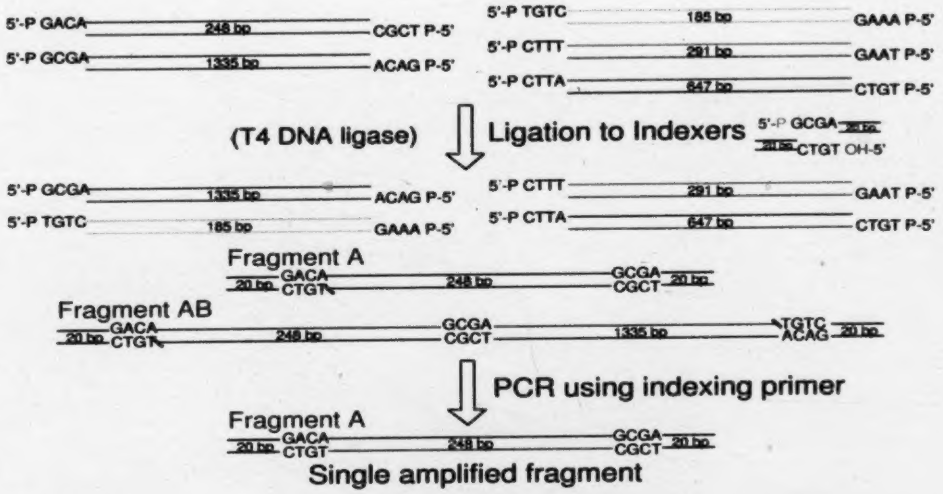
The BamCC indexers are named for the 2-nt CC extension on the 3'-end of the BamCC sequencing primer. The 26-nt BamCC indexing strand (FIGURE 2.10B) remained unphosphorylated following synthesis, and bore a sequence of the general format 5'-OH-NNNNGGTTGTCTTCGGATCCTGTACC-3'. Any particular BamCC indexing strand oligo thus bore a nonphosphorylated informative 5' 4-nt cohesive end corresponding to one of the 256 tetranucleotide sequences, a two-nt GG extension (used to provide directionality of template priming for cycle sequencing) inserted between the cohesive end and the primer-binding sequence, and the 20-nt Bam-primer-binding sequence (5'-TTGTCTTCGGATCCTGTACC-3') common to all indexing strands of both the Bam and BamCC indexer sets. The 22-nt BamCC sequencing primer strand (5'-OH-GGTACAGGATCCGAAGACAACC-3') was annealed to each BamCC indexing strand to generate a set of double-stranded nonphosphorylated indexers with 4-nt cohesive ends.

2.3.10 Use of the Bam and BamCC indexer sets for compound-primer P/NoP indexing

The Bam and BamCC indexer sets are designed to be used in conjunction with each other to target and amplify indexable fragment classes in a manner employing a single primer sequence for amplification while enabling subsequent end-specific priming of cycle sequencing. The *compound-primer P/NoP indexing* strategy thus incorporates the P/NoP approach preventing the amplification of repeated-end fragments, provides the means for directional cycle sequencing of amplified indexed fragments, and simultaneously eliminating PD artifact amplification from indexing PCR reactions. In an indexing reaction employing this strategy, one phosphorylated standard (Bam) indexer and one nonphosphorylated compound (BamCC) indexer bearing the same common-primer-binding region are ligated to the target fragment. Amplification proceeds using the single 20-nt common primer, thus avoiding the production of primer-dimer. Subsequent to amplification, the target fragment may be sequenced from the compound-primer end alone by using the 22-nt CC-extended primer specific for the compound primer.

FIGURE 2.11 outlines the process by which the compound-primer P/NoP strategy permits amplification of a target fragment using a common primer to the exclusion of non-targeted repeated-end fragments, and subsequently enables the unidirectional priming of DNA strand elongation for cycle sequencing of the amplified indexed template. Fragment A is used as an example. Indexers P-GCGAxBam and OH-TGTCxBamCC are added to a ligation reaction containing *Fok*I-digested pUC19 DNA and T4 DNA ligase. Accurate joining of indexers to their complementary cohesive end sequences permits the ligation of both indexing and primer strands of P-GCGAxBam to the TCGC cohesive end of Fragment A. The 5'-phosphate on the GACA cohesive end sequence of Fragment A is joined to the 3'-hydroxyl of the primer strand of OH-TGTCxBamCC, but ligation of the indexing strand to the target is prevented by the lack of a 5'-phosphate on the indexing strand. Ligation of OH-TGTCxBamCC to the GACA cohesive end of Fragment B likewise occurs only between the target and the primer strand of the indexer. The legitimate religation of the TCGC end of Fragment A with Fragment B's GCGA end sequence to reassemble the AB double-fragment occurs uninhibited, and some AB molecules will bear semi-ligated OH-TGTCxBamCC on both ends.

Upon denaturation of indexed Fragment A during the first cycle of PCR, the unligated BamCC indexing strand will dissociate from the target and its Bam-primer-binding site will be lost. However, priming of DNA strand elongation from the other


5'-P GACA
248 bp
CGCT P-5'
5'-P GCGA
1335 bp
ACAG P-5'
5'-P TGTC
185 bp
GAAA P-5'
5'-P CTTT
291 bp
GAAT P-5'
5'-P CTTA
647 bp
CTGT P-5'
(T4 DNA ligase)
Ligation to Indexers
5'-P GCGA
20 bp
20 bp
CTGT OH-5'
5'-P GCGA
1335 bp
ACAG P-5'
5'-P TGTC
185 bp
GAAA P-5'
5'-P CTTT
291 bp
GAAT P-5'
5'-P CTTA
647 bp
CTGT P-5'
Fragment A
20 bp
GACA
CTGT
248 bp
GCGA
CGCT
20 bp
Fragment AB
20 bp
GACA
CTGT
248 bp
GCGA
CGCT
1335 bp
TGTC
ACAG
20 bp
PCR using indexing primer
Fragment A
20 bp
GACA
CTGT
248 bp
GCGA
CGCT
20 bp
Single amplified fragment

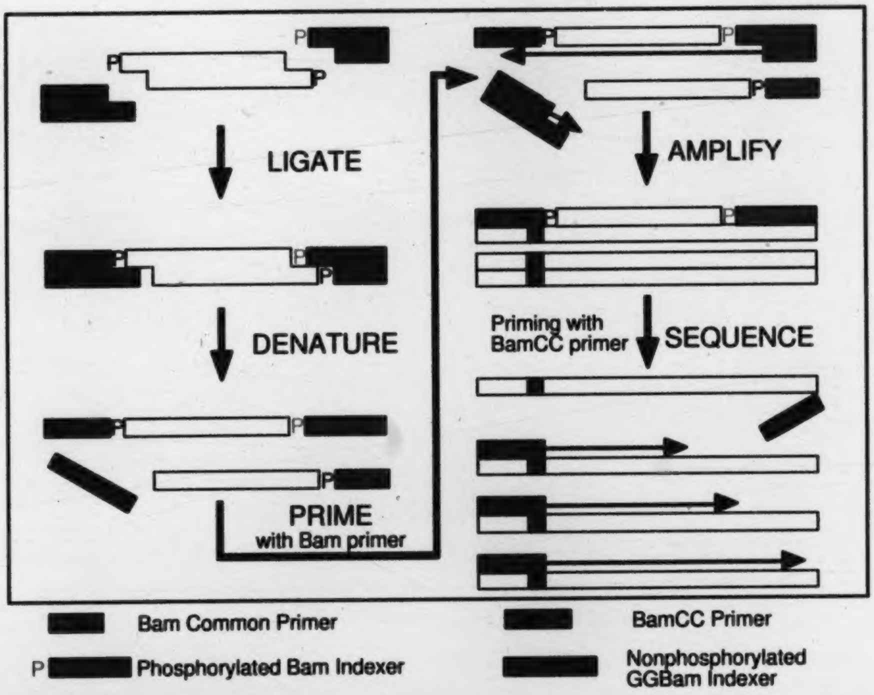

P
P
P
LIGATE
P
P
P
DENATURE
P
P
P
PRIME
with Bam primer
P
P
P
AMPLIFY
P
P
Priming with
BamCC primer
SEQUENCE
Bam Common Primer
P
Phosphorylated Bam Indexer
BamCC Primer
Nonphosphorylated
GGBam Indexer

FIGURE 2.11: Compound-primer P/NoP strategy.
The compound-primer P/NoP strategy permits amplification of a target fragment using a common primer without amplifying non-targeted repeated-end fragments, and enables unidirectional template priming for sequencing.See text for details.

end of the molecule by a Bam primer binding to its complementary sequence on the ligated Bam indexing strand results in the reconstitution of a complete second strand bearing an intact Bam-primer-binding site. Thus, although the first cycle of PCR results in only linear amplification of the target fragment, amplification during each subsequent cycle will be exponential. In contrast, both the (unligated) Bam-primer-binding sites dissociate from the AB repeated-end fragment, and amplification fails. Therefore Fragment A, the desired indexing target, is the only amplified product in the reaction (as the PD artifact is not produced from a single-primer PCR), and is ready for use as a sequencing template directly from the amplification reaction. (Amplification of the repeated-end CDE triple fragment is prevented in a similar manner in ligations containing the OH-GACAxBamCC indexer.)

Cycle sequencing is primed by the BamCC directional sequencing primer. As a complete BamCC binding site is only found at one end of the amplified indexed fragment, initiation of strand elongation will proceed only from that end. The BamCC indexers have a 22 nt core sequence, the first 20 nt of which is common to both the Bam and BamCC indexer sets. Nonetheless, the 2 bases at the 3' end of the BamCC core sequence are sufficient to inhibit strand elongation from a BamCC primer that has erroneously annealed to the Bam-primer-binding sequence at the opposite end of the fragment, due to *Taq* DNA polymerase's capacity to discriminate against a 3'-nucleotide mismatch successfully inhibits strand elongation [129]. *Taq* DNA polymerase is highly specific for template complementarity for primers with GG as the 3' dinucleotide, providing substantial template discrimination (40- to 100-fold) against 3'-end mismatches. Longer unique sequences could be used but 2 nt of unique 3' sequence is sufficient to discriminate between sequencing primer sites.

An additional indexer set, the phosphorylated BamGG set, was designed that would allow sequencing to be performed from one end, then the other, of a template from the same amplification reaction. In the context of the simple pUC19 model system, the Bam/BamCC approach was deemed to be a sufficient demonstration of the principle. In order to obtain sequence data in both directions along the fragment, the "polarity" of the indexers used in the reaction was reversed. For example, following the indexing and amplification of Fragment A by P-GCGAxBam and OH-TGTCxBamCC to permit sequencing from the "left" end of the fragment, Fragment A was indexed using OH-GCGAxBamCC and P-TGTCxBam, permitting sequencing from the "right" end of the fragment. Even though the use of the repeated-end indexer will permit legitimate amplification of the AB double-fragment, Fragment A can still be sequenced directly out of the amplification reaction without "sequence crosstalk" with AB, as AB does not carry a BamCC directional sequencing primer-binding site.

2.3.11 Compound-primer P/NoP indexing of *Fok*I-digested pUC19 fragments

The compound-primer P/NoP indexing of pUC19 by fragments of the Bam P-indexer set and the BamCC NoP indexer set was performed according to protocols previously demonstrated to efficiently amplify the pUC19 target fragments when a single set of indexers were used. Fragment A was indexed with P-GCGAxBam and OH-TGTCxBamCC; Fragment B by P-TCGCxBam and OH-TGTCxBamCC; Fragment C by P-CTTTxBam and OH-GACAxBamCC; Fragment D by P-AAAGxBam and OH-CTTAxBamCC; and Fragment E by P-TAAGxBam and OH-GACAxBamCC. All ligations contained 1 ng of *Fok*I-digested pUC19 DNA and 40 U of T4 DNA ligase, were incubated for 1 h at 37°C, and halted by denaturation of the ligase at 65°C for 15 min. Two microlitres of each completed ligation reaction was added to a 50-µl PCR reaction containing antibody-sequestered *Taq* DNA polymerase and utilizing 40 pmol of Bam primer as the sole indexing primer. Following a hot start, amplification proceeded for thirty cycles under standard indexing PCR conditions. Amplification products were analyzed by agarose gel electrophoresis (FIGURE 2.12).

The effective use of the Bam/BamCC indexing sets to perform compound-primer P/NoP indexing was demonstrated. Each of the pUC19 target fragments was efficiently indexed and amplified. Indexing reaction protocols were developed that provided high ligation fidelity without sacrificing ligation efficiency. No misligation products were amplified. The use of the P/NoP strategy circumvented the amplification of repeated-end fragments. The compound-primer strategy permitted the

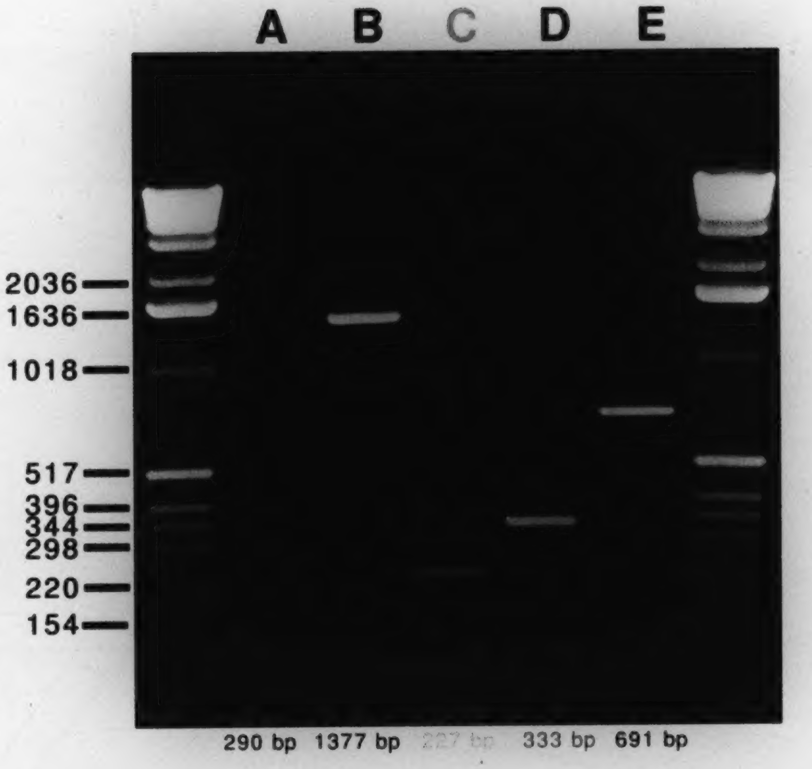


FIGURE 2.12: Compound-primer P/NoP indexing of pUC19 fragments. Fragment A was indexed with **P**-GCGAxBam and **OH**-TGTCxBamCC; Fragment B by **P**-TCGCxBam and **OH**-TGTCxBamCC; Fragment C by **P**-CTTTxBam and **OH**-GACAxBamCC; Fragment D by **P**-AAAGxBam and **OH**-CTTAxBamCC; and Fragment E by **P**-TAAGxBam and **OH**-GACAxBamCC. All ligations contained 1 ng of *Fok*I-digested pUC19 DNA and 40 U of T4 DNA ligase, were incubated for 1 h at 37°C, and halted by denaturation of the ligase at 65°C for 15 min. Two microlitres of each ligation was added to a 50-ml PCR reaction containing antibody-sequestered *Taq* DNA polymerase and 40 pmol of Bam primer as the sole indexing primer. Following a hot start, amplification proceeded for 30 cycles. Amplification products were analyzed by agarose gel electrophoresis.

amplification of target fragments in a form that is directly and directionally sequenceable, without the production of PD artifacts.

2.3.12 Indexing of pUC19 *Fok*I fragments for sequencing template production

Each of the five pUC19 *Fok*I fragments was indexed and amplified twice using the Bam/BamCC compound-primer P/NoP approach. In one set of indexing ligations, the BamCC indexer targets and is ligated to the cohesive end sequence at the "left" end of the restriction fragment (i.e. the 5'-end of the fragment if the standard nucleotide numbering convention employed in the GenBank single-strand pUC19 sequence is used [74]). This imports the BamCC primer-binding sequence onto the left end of the target fragment, providing the means for use of the amplified indexed fragment as a template for directional sequencing initiated at the left end of the molecule. In the other set of indexing reactions, the BamCC indexer ligates to the cohesive end sequence at the "right" end of the target fragment, permitting amplification of a sequencing template with the opposite directionality. In addition, Bam indexers were used for the P/P indexing of each pUC19 *Fok*I fragment, providing specific size markers for each of the P/NoP amplicons and highlighting the ability of appropriately-selected P/NoP indexers to reduce the contribution of repeated-end fragment class amplification to indexing analyses complexity.

Amplification products from these three sets of indexing reactions were analyzed by agarose gel electrophoresis (FIGURE 2.13). The slight difference in length between P/P-indexed fragments and the related fragments indexed using the P/NoP approach is due to the two extra base pairs present in the BamCC indexer sequence used in P/NoP indexing. The AB repeated-end fragment is amplified from ligations targeting either Fragment A or Fragment B which contained the TGTC P-indexer. As discussed in Section 2.3.10, even target fragments amplified in the legitimate presence of repeated-end fragments, due to the choice of phosphorylated Bam indexer sequence in the ligation, may still be sequenced directly from the PCR reaction without crosstalk from the other amplification products present. As the repeated-end fragment bears only the Bam-primer-binding sequence at both ends,

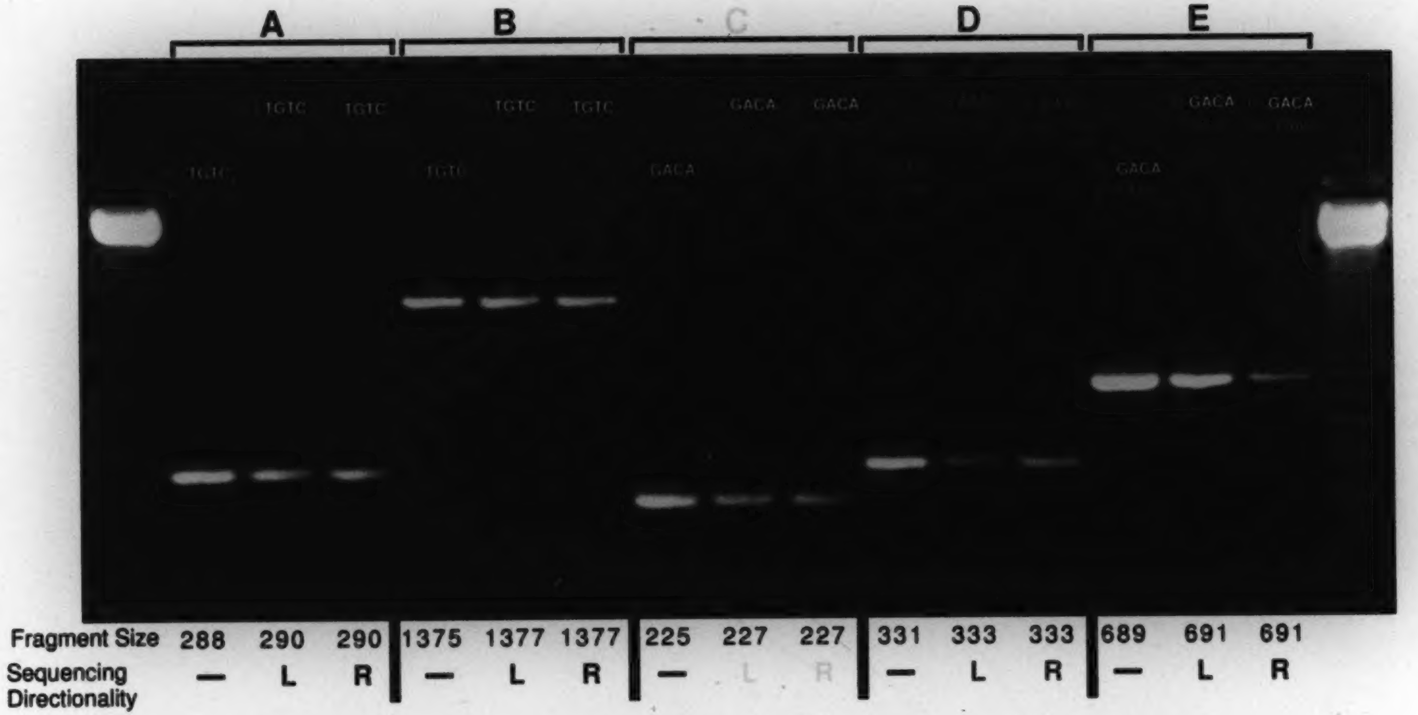

A
B
C
D
E
TGTC
GACA
Fragment Size 288 290 290 1375 1377 1377 225 227 227 331 333 333 689 691 691
Sequencing Directionality — L R — L R — L R — L R — L R

FIGURE 2.13: Indexing of pUC19 *FokI* fragments for sequencing template production.
Each pUC19 *FokI* fragment was indexed and amplified twice using the compound primer P/NoP approach, once with the NoP indexer targeting the "left" cohesive end of the fragment and once with the NoP indexer targeting the "right" cohesive end. In addition, each fragment was indexed and amplified using the P/P indexing strategy providing site-specific markers for each pair of P/NoP amplicons. Ligation, PCR and agarose gel electrophoresis were performed as previously described.

neither end of the molecule contains a BamCC sequencing-primer-binding site. Only the target fragment, generated by ligation to both a Bam indexer and to a BamCC indexer, is amenable to DNA strand elongation from an accurately base-paired BamCC primer during cycle sequencing.

The CDE amplicon, a legitimate indexing target in ligations of Fragments C or E containing the GACA P-indexer, is not detected in these amplifications. This may be due to the lower specific amplification efficiency of the completely-assembled CDE amplicon relative to the smaller amplicons present in the same amplification reactions. Due to the lower number of consecutive nucleotide chain elongation events required for full extension of the replicating strand of DNA across the primer-binding sequence at the opposite end of the template, or possibly due to the effect of %GC differences of the fragment sequences between the indexer sequences on dissociation rates during the melting and annealing phases of PCR, amplification of the smaller amplicons outcompetes that of the large CDE fragment. Alternately, the reaction kinetics involved in the assembly of an indexed CDE amplicon during ligation may affect its ability to be amplified in a competitive PCR reaction.

Unlike the assembly of a single-fragment indexed amplicon which is the result of two independent iterations of two successive second-order reactions (see Section 2.1.3), a re-ligated CDE fragment with indexers ligated to both ends is assembled only following five independent iterations (indexer 1 + Fragment C + Fragment D + Fragment E + indexer 2) of two successive second-order reactions. In addition, Bam/BamCC indexing of pUC19 fragments for the generation of sequencing templates took place under much less permissive (and thus less efficient) ligation conditions than those employed in indexing experiments in which the CDE amplicon was observed (compare FIGURE 2.6). The low rate of CDE amplicon assembly efficiency during ligation results in a much lower copy number of CDE initially present in the amplification reaction relative to its competitors. The copy number discrepancy between indexed CDE and indexed single-fragment targets exacerbates the effect of CDE's lower amplification efficiency. Both of these factors may contribute to the failure of the CDE triple-fragment to be amplified in a competitive PCR to sufficient levels for visualization on an agarose gel.

2.3.13 Amplification of gap-closing B_a and B_b sequencing templates from a *Fok*I/*Sfa*NI double restriction digest

For digests of DNA with 50% G/C content, the average size of a *Fok*I fragment is expected to be about 512 bp. On average, the number of *Fok*I fragments in a complex digest (e.g. of a 5 Mb bacterial genome) with a length less than 1250 bp is likely to be more than 92.5% of the total number of different fragments in the digest. This size range is appropriate for fragment amplification and for direct cycle sequencing. The length of a typical sequence read obtained from fluorescently-labeled dideoxynucleotide sequencing data analyzed using an Applied Biosystems 377 slab gel automated DNA sequencer is about 500 bp [130]. Each of the *Fok*I restriction fragments of pUC19 is easily amplified using indexing primer oligos following ligation of indexers to both ends of each fragment. In the cases of Fragments A, C, D and E, the resulting indexing amplicons are of a size amenable for cycle sequencing across the template in both directions, often obtaining complete template coverage in any one direction.

Fragment B (1335 bp prior to ligation of indexers) is of a size that is representative of about 5% of the fragments found in a typical *Fok*I digest of a 5 Mb bacterial genome. Amplification of the indexed Fragment B amplicon is straightforward and requires no special measures. However, sequencing of indexed Fragment B is problematic due to the length of the template and the resultant distance to covered by sequencing initiated from primers at either end of the molecule. Use of the BamCC directional sequencing primer from the left side and subsequently the right side of Fragment B typically leaves about 300 bp in the middle of the molecule uncovered by typical sequence read lengths. While only a small percentage of *Fok*I fragments in a digest will be refractory to full-coverage sequencing, solutions facilitating full coverage developed for Fragment B may be applied to real-world instances of this challenge to indexing-based sequencing efforts. The inclusion of a

*Fok*I fragment presenting challenges mimicking those of real-world systems is a characteristic contributing to the utility of the pUC19 model system.

A simple solution to the challenges involved in indexing-based sequencing of large templates is the use of a second Type IIS enzyme to subdivide the refractory fragment, permitting indexing, amplification and directional sequencing of the subfragments. The Type IIS restriction endonuclease *Sfa*NI has a recognition sequence (5'-GCATC(N)$_{5/9}$-3') that, like *Fok*I, is expected to occur once every 512 bp on average in 50% GC DNA. The cohesive ends generated by cleavage by *Sfa*NI are similar to those produced by *Fok*I: 4-nt 5' quasi-random cohesive ends that are specific for any cleavage location but unpredictable from the enzyme's recognition sequence. The sequence of pUC19 contains 8 *Sfa*NI cleavage sites. Between the two *Fok*I sites that generate Fragment B from intact pUC19, there is a single *Sfa*NI site, cleaving Fragment B (1335 bp) into two indexable fragments (FIGURE 2.14). Fragment B$_a$, derived from the left half of Fragment B, is 573 bp in length and bears the cohesive end sequences GCGA (from the left-side *Fok*I cleavage site) and GATT (from *Sfa*NI cleavage). Fragment B$_b$, derived from the right half of Fragment B, is 766 bp long and presents the cohesive end sequences AATC (from *Sfa*NI cleavage) and GACA (generated by *Fok*I cleavage on the right side of the fragment). These two *Fok*I/*Sfa*NI fragments are easily indexed directly from a *Fok*I/*Sfa*NI double restriction digest of pUC19 using the Bam/BamCC indexer sets (FIGURE 2.14). If the *Sfa*NI-cut cohesive end of each fragment is targeted by the BamCC indexer in each ligation, the amplified indexed fragments may be used as templates for cycle sequencing of the central region of Fragment B previously refractory to sequencing. Using the directional sequencing primer BamCC, sequencing proceeds from the middle of the unsequenced region outward along both the B$_a$ and B$_b$ templates towards the regions of Fragment B sequenceable directly from Fragment B templates, providing overlapping coverage and gap closure. This solution has general applicability to *Fok*I restriction fragments of a length refractory to complete sequence coverage from end-initiated sequencing alone. Any *Fok*I fragment of such a length is likely to contain interior *Sfa*NI recognition sites that may be used to generate indexable, amplifiable and

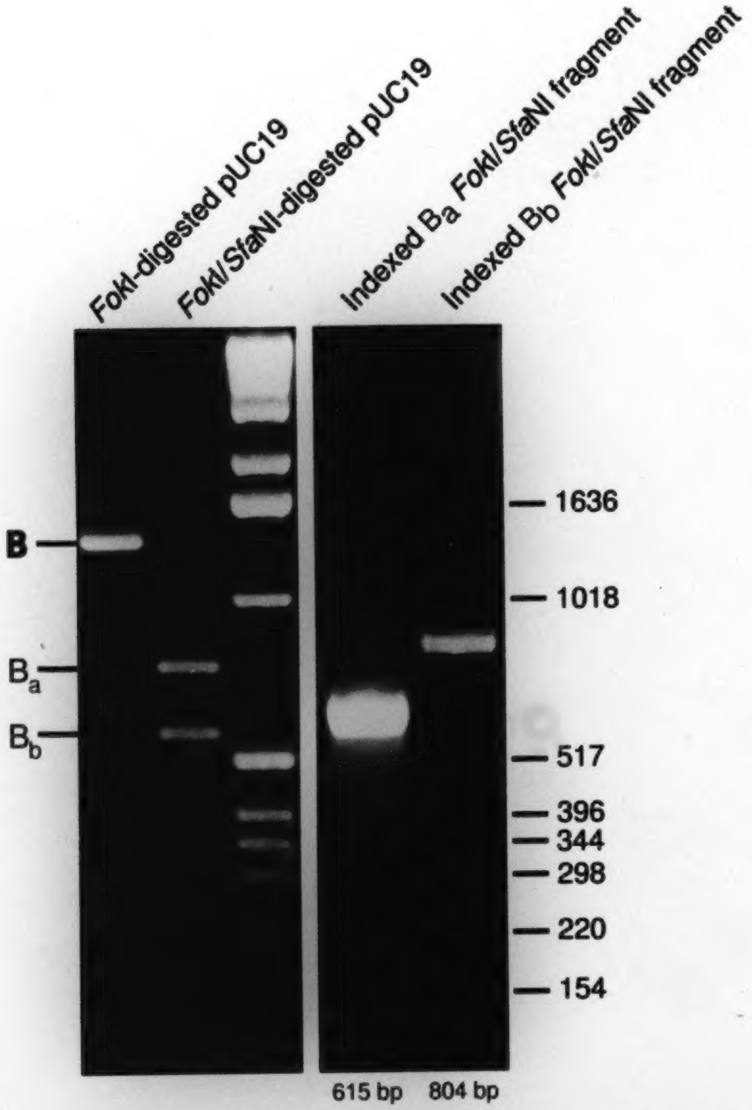

FokI-digested pUC19
FokI/SfaNI-digested pUC19
Indexed Ba FokI/SfaNI fragment
Indexed Bb FokI/SfaNI fragment
B
Ba
Bb
1636
1018
517
396
344
298
220
154
615 bp
804 bp

FIGURE 2.14: Amplification of B_a and B_b sequencing templates from a *Fok*I/*Sfa*NI double restriction digest.

Cleavage of *Fok*I-digested pUC19 DNA by *Sfa*NI endonuclease generates a 573-bp *Fok*I/*Sfa*NI fragment (B_a) and a 766-bp *Fok*I/*Sfa*NI fragment (B_b), among others. Digestion of pUC19 DNA and subsequent analysis of the digested DNA by agarose gel electrophoresis were performed as described in **Section 2.2.10.** A ligation reaction containing 1 ng *Fok*I/*Sfa*NI-digested pUC19 DNA and indexers P-TCGCxBam and OH-AATNxBamCC was assembled using indexers P-TGTCxBam and OH-GATNxBamCC. Ligation, PCR and agarose gel electrophoresis were performed as previously described.

sequenceable subfragments, providing gap closure. Existing sets of indexers may be used to accomplish these tasks.

2.3.14 Direct cycle sequencing of amplified indexed pUC19 fragments

Two indexed sequencing templates assembled from each pUC19 FokI fragment by ligation of Bam/BamCC indexers with opposite sequencing directionality (see FIGURE 2.13*)*, and the two *Fok*I/*Sfa*NI gap-closing sequencing templates B_a and B_b (see FIGURE 2.14) were amplified using *PfuTurbo*™ DNA polymerase and purified as described (see Section 2.2.11). Following amplification, cycle sequencing of the twelve sequencing templates was performed using a commercial fluorescently-labeled dideoxynucleotide cycle sequencing kit. The DNA indexing sequencing primer BamCC was used to initiate sequencing only from the end of the indexed fragments to which the non-phosphorylated BamCC indexer had been ligated. This unidirectionality ensured that the cycle sequencing reactions were not contaminated by simultaneous initiation of sequencing from both ends of the indexed template. Multiple iterations of amplification, cycle sequencing and data analysis by automated sequencing instrumentation were performed in order to provide redundant coverage of sequencing templates. Sequence data was assembled and processed using commercial contig assembly software.

2.3.15 Alignment of indexing-based directionally-sequenced pUC19 templates to an indexing-based restriction map of pUC19 constructed by jigsaw assembly

Assembly of restriction-based physical maps of DNAs digested with Type IIS and IP endonucleases is simplified by knowledge of the cohesive end sequences of each restriction fragment used in map construction. The resolution of the map is determined by the average fragment length generated by the restriction endonuclease used for mapping, a consequence of the frequency of occurrence of the enzyme's recognition sequence. *Fok*I endonuclease cleaves DNA every 512 bp on average, and may be used to generate high resolution maps of small DNA molecules such as pUC19.

As the contiguous cohesive end sequences of adjacent fragments must be complementary, ordering of fragments along the map may be directly determined by matching consecutive complementary-sequence-bearing fragments from the pool of known end sequences. This simple approach to physical map construction is referred to as *jigsaw assembly*.

Appropriate selection of restriction enzyme and indexing strategy permits subdivision of the DNA to be mapped such that the cohesive end sequence information uniquely defines the order of fragments along the physical map, and no additional measures are required to determine a unique mapping solution. For instance, a single map is expected from *Fok*I-based jigsaw mapping of DNAs smaller than 50 kb. For larger DNAs, the presence of a particular end sequence on multiple fragments in the Type IIS restriction digest may produce multiple putative mapping solutions which can be resolved using other techniques (e.g. jigsaw mapping with a second Type IIS enzyme, Southern blotting, fluorescent *in situ* hybridization, etc.). The number of possible linkages between duplicated ends, and thus the number of alternative map constructions, is severely constrained by the requirement for map closure. Preliminary calculations indicate that in a random population of 100 fragments bearing 4-nt cohesive end sequences, 6 to 8 cohesive end sequences are expected to be represented more than once. (Some 4-nt cohesive end sequences are palindromic, and are thus complementary to themselves. Cleavage at a site that produces these types of ends will necessarily generate two fragments bearing the same end sequence on one of each of their ends.) Inspection of test cases with 8 duplicated end sequences resulted in generation of only two putative mapping solutions per case [131]. In digests in which the same cohesive end sequence occurs more than four times, the number of possible maps increases factorially. In such cases, each alternate map may be treated as a testable hypothesis for map assembly that may be proven or rejected on the basis of easily-obtained additional information. Alternately, selection of different mapping enzymes or indexing strategies may facilitate the assembly of a unique physical map by jigsaw assembly. For instance, indexing-based restriction mapping with *Hga*I (which produces 5-nt cohesive ends and therefore provides a larger number of cohesive end sequence classes than *Fok*I) and therefore greater specificity for jigsaw assembly of indexing-based restriction maps by reducing the probability of duplicated ends.

The sequence data analysis and contiguous sequence (*contig*) assembly software Sequencher v3.0 was used to construct a index map (FIGURE 2.15) and a completely sequenced contig (FIGURE 2.16) of the pUC19 vector. Automated sequence data of indexed pUC19 DNA fragments in raw chromatogram form were trimmed of indexer sequence and manually edited for obvious miscalled bases. Trimmed sequences were compared, matched and aligned by the software's contig assembly algorithm into subcontigs corresponding to each of the indexed pUC19 fragments used as templates for sequencing. Subcontigs were linked by their cohesive end sequences by comparing the subcontigs to the index map to assemble the wholely sequenced contig of pUC19. In this demonstration of DNA indexing concepts, the compound-primer strategy was successfully applied to the pUC19 model system in a manner that permitted the complete characterization and sequencing of the system.

2.4 DISCUSSION

The pUC19 model system was successfully exploited in the development of indexing strategies and appropriate indexing protocols. Model system characteristics presenting challenges to indexing approaches mimicking those presented in more complex systems were investigated, and indexing strategies developed to meet those challenges. The compound-primer strategy was successfully applied to the pUC19 model system, in a manner that permitted the complete characterization and sequencing of the system.

On the basis of these results and the core principles involved in DNA indexing approaches, a proposal for a future application of DNA indexing is presented (see **Section 5.2**). A model for the use of multiplex indexing in the identification of Type IIS cohesive end sequences is described. The use of indexing sequence tagged sites, generated from pairs of indexed S/F fragments centered on *Sfi*I restriction sites, for physical mapping of complex genomes is outlined. Finally, an entirely indexing-based method for directed mapping and sequencing of prokaryotic genomes is proposed.

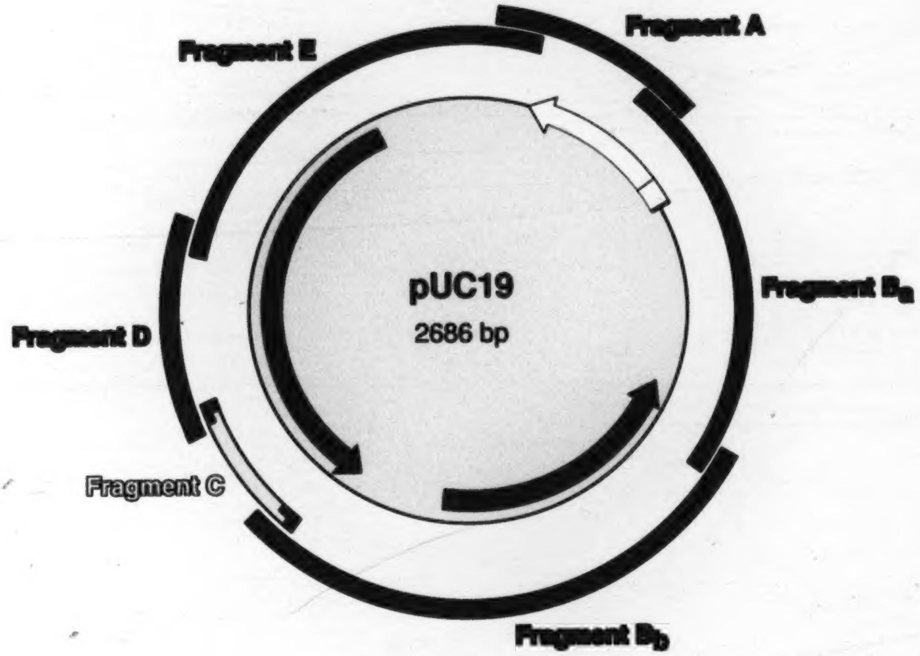

Fragment A
Fragment E
Fragment B_a
pUC19
2686 bp
Fragment D
Fragment C
Fragment B_b

FIGURE 2.15: Index map of pUC19 sequencing template coverage. See text for details.

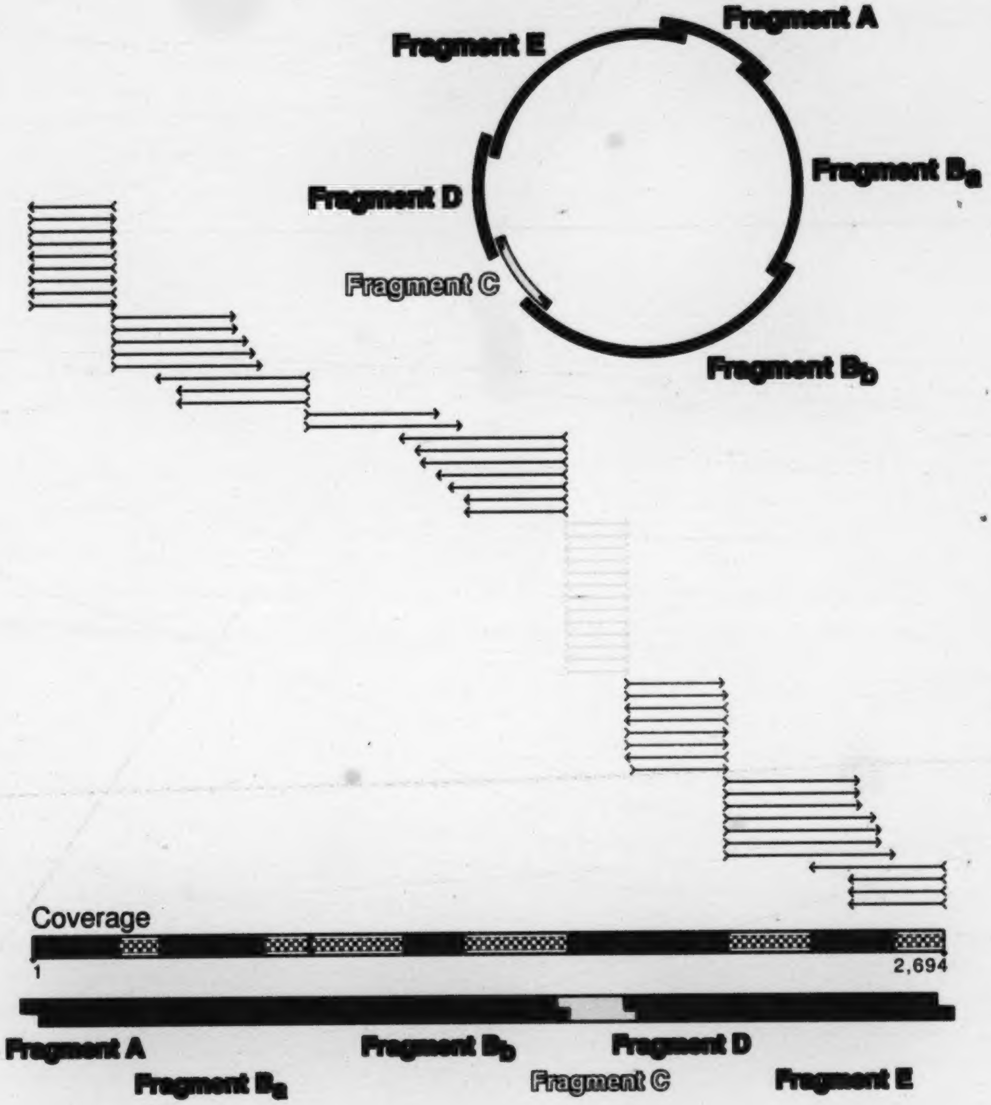

Fragment A
Fragment E
Fragment B_a
Fragment D
Fragment C
Fragment B_b
Coverage
1
2,694
Fragment A
Fragment B_b
Fragment D
Fragment B_a
Fragment C
Fragment E

FIGURE 2.16: Complete indexing-based sequencing contig of pUC19. Coverage of pUC19 sequence from both directions is indicated in green. Redundant coverage obtained by sequencing from a singel directions is indicated in blue checks. See text for other details.

3 Chapter III: Bacterial Strain and Species Differentiation by Indexed Genomic Profiling (IGP)

3.1 INTRODUCTION

The ability to distinguish between similar bacterial species and between closely-related bacterial strains is essential for the clinical microbiologist and the microbial epidemiologist. The emergence of antibiotic-resistant strains of several endemic bacterial species magnifies the importance of effective tools for microbial typing for the investigation of infection clusters within hospitals or communities. Typing methods can be broadly categorized as focusing on either the phenotype or the genotype of the isolates under investigation.

Historically, prokaryotes have been classified on a phenotypic basis, characterizing bacterial species and strains using such criteria as morphology, biochemical or metabolic profile, serotype, or antibiotic susceptibility. While phenotypic methods of bacterial typing are typically inexpensive and simple to perform in a clinical laboratory setting, in some cases these approaches are incapable to resolve closely-related bacterial strains [132]. Some phenotypic typing systems developed for classifying clinical isolates of specific species are not generally applicable due to the specially adapted reagents employed [133]. Other phenotypic methods are subject to the variation associated with studies of gene expression, resulting in poor reproducibility for these approaches [134]. Aggravating this potential for irreproducibility is a lack of both standardized methodology and objective standards for data evaluation for many applications [135]. To circumvent the deficiencies inherent in phenotypic methods of bacterial typing, significant effort has been focused on the development of a range of genotypic techniques capable of wide applicability, excellent reproducibility and increased discriminatory power [134].

Genotypic approaches to bacterial species and strain typing, based on the genetic variation present in the chromosomal DNA of a bacterial species [136], have several advantages over traditional phenotypic methods. They are typically more broadly applicable to a variety of bacterial species, present higher discriminatory power, and are in certain instances more time-efficient [137]. Many approaches to bacterial genotyping have been employed in both research and clinical laboratory settings. These include restriction fragment length polymorphism (RFLP) analysis [50], ribotyping [138], pulsed-field gel electrophoresis (PFGE) genome analysis [139], PCR-based methods such as arbitrarily-primed PCR (AP-PCR) [140, 141] and random amplification of polymorphic DNA (RAPD) [141], and amplified-fragment length polymorphism (AFLP) [142-144].

The "gold standard" for evaluation of genotypic typing techniques has been characterized as a technique providing "optimal typeability, a high degree of reproducibility, adequate stability, and unprecedented resolving power" [135]. While each molecular approach offers advantages over the various phenotypic methods for strain typing, there is also opportunity for improvement. The utility of RFLP analysis is highly dependant on appropriate oligonucleotide probe choice, is applicable only to strains for which probes have been specifically designed, is subject to inaccuracy due to partial digestion or poor resolution, and does not provide unambiguous interpretation of results in all cases [134, 145]. Although ribotyping and PFGE have been demonstrated to provide adequate discriminatory power for typing various bacterial pathogens [135, 146], these methods are time-consuming and may inherently lack discriminatory power [147-149]. RAPD analysis and AP-PCR, while appropriate for intralaboratory comparisons of microbial clinical isolates prior to stringent objective analysis, do not provide the reproducibility necessary for fully descriptive genotyping or interlaboratory comparisons [83, 135]. Although the selective amplification of a subset of restriction fragments from a genomic digest by AFLP to differentiate effectively between closely related strains for a wide range of bacterial species [144, 149, 150], it does not offer the information density, cost-effectiveness, or the exquisite discriminatory power available to an indexing-based approach to bacterial genotyping for species and strain discrimination.

3.1.1 ABSTRACT:

An indexing-based approach to microbial molecular subtyping was developed and demonstrated. Existing indexing protocols were adapted for the complexity of microbial genome analysis and to provide increased information density in experimental data. Initial application of the modified protocols to the molecular fingerprinting and differentiation of several *E. coli* strains was accompanied by predictive modeling based on the published genomic DNA sequence of *E. coli* strain MG1655. Indexed genomic profiles were generated from clinical isolates and reference strains of several *Staphylococcus* species. Indexed genomic profiling (IGP) provides excellent discriminatory power in the form of an information-dense molecular fingerprint derived by objective sampling of microbial genetic structure. Its specificity, widespread applicability, reproducibility, and potential for high-throughput application make IGP an attractive method for microbial typing in clinical or research laboratory environments.

3.2 MATERIALS AND METHODS

3.2.1 Bacterial strains used in IGP investigations

Escherichia coli strains employed in this investigation were purchased as lyophilized cultures from the American Type Culture Collection (ATCC; Rockville ML). *E. coli* MG1655 [F^- λ^-] (ATCC 47076) is a well-characterized transformation host strain derived from "wild-type" strain K12 that does not carry the F plasmid and for which the genomic DNA sequence is known [151]. *E. coli* JM109 [*recA1, endA1, gyr96A, hsdR17, thi, relA1, supE44*, Δ*(lac-proAB)* (F', *traD36, proAB, lac*I^{q}ZΔM15), λ^-] (ATCC 53323) is a recombination-deficient strain commonly used as a host for site-directed mutagenesis or expression [74, 152]. *E. coli* JM110 [*dam, dcm, rpsL, thr, leu, thi, lacY, galK, galT, ara, tonA, tsx, supE44*, Δ*(lac-proAB)* (F', *traD36, proAB, lac*I^{q}ZΔM15) (λ-)] (ATCC 47013) is a methylation-deficient strain that is a derivation of strain JM109 [153]. *E. coli* W3110 [IN(*rrnD-rrnE*) F^{-} λ^{-}] (ATCC 27325) is a strain closely related to strain MG1655 for which the genome sequence has also been determined, although genomic sequencing was performed using a methodology thought to be more prone to inaccuracies than that used for the sequencing of strain MG1655 [151, 154]. The cultures were rehydrated according to the supplier's instructions and cultured using standard laboratory procedures.

The genome of the laboratory standard strain *Staphylococcus aureus* subsp. *aureus* NCTC 8325-4 (ATCC 12600) is currently being sequenced by the *Staphylococcus aureus* Genome Sequencing Project in the University of Oklahoma's Department of Chemistry and Biochemistry (Norman, Oklahoma OK). Cultures of *S. aureus, S. epidermidis* 9759 (ATCC 14990) and *S. lugdunensis* (ATCC 43809) were kindly donated by Dr. Anthony Chow (Faculty of Medicine, University of British Columbia). *Staphylococcus* species were maintained and cultured in accordance with standard laboratory practice.

3.2.2 Isolation and purification of bacterial genomic DNA

Genomic DNA was isolated from *E. coli* and *Staphylococcus* cultures using a Genomic-tip DNA Purification system (QIAgen). Cell lysis and DNA purification from *E. coli* cultures proceeded as outlined by the manufacturer's recommended protocol. *Staphylococcus* cultures were lysed by the addition of 2 mg of lysostaphin (Sigma-Aldrich) in addition to the manufacturer's suggested amounts of protease and lysozyme. DNA purification was then performed in a manner similar to that employed with the *E. coli* DNA purifications.

3.2.3 Digestion of bacterial DNA with *Fok*I restriction endonuclease

Restriction digests of bacterial genomic DNAs were performed in a manner similar to that employed in the digestion of pUC19 with *Fok*I restriction endonuclease, using 1 μg of bacterial genomic DNA in each digest.

3.2.4 Synthesis and annealing of indexing oligonucleotides

The phosphorylated Bam indexing oligos, nonphosphorylated BamCC indexing oligo mixes, Bam primers and BamCC primers used in the mapping and sequencing of pUC19 were synthesized, annealed and stored in the manner described in **Section 2.3.9**.

3.2.5 Ligation of indexers to *Fok*I-digested bacterial genomic DNA: standard conditions

Typical indexing ligation reactions of bacterial genomic DNA contained 10 ng of *Fok*I-digested bacterial DNA, 50 fmol of the appropriate P-indexer, and 50 fmol/indexer of the appropriate NoP indexer mix (200 fmol total NoP indexer) per ligation. Ligations employing T4 DNA ligase were prepared and performed using the ligation buffers, incubation temperatures and incubation times described in **Chapter II**, unless otherwise indicated.

Ligations containing *Taq* DNA ligase (New England Biolabs) were assembled in 20-µl reaction volumes of NEB *Taq* DNA Ligase Buffer [20 mM Tris-HCl, 25 mM $C_2H_3KO_2$, 10 mM $Mg(CH_3COO)_2$, 10 mM DTT, 1 mM NAD, 0.1% Triton X-100, pH 7.6 @ 25°C], incubated for 1 h at 42°C, and terminated by cooling the reaction to 0°C. All other reaction constituents were added in the amounts and concentrations specified in the figures.

Ligations containing *E. coli* DNA ligase (New England Biolabs), 10 ng of *Fok*I-digested *E. coli* DNA, 50 fmol of the specified P-indexer and 200 fmol [total indexer] of the specified NoP indexer mix, were assembled in 20-µl reaction volumes of NEB *E. coli* DNA Ligase Buffer [50 mM Tris-HCl, 10 mM $MgCl_2$, 10 mM DTT, 26 mM NAD, 25 ng/µl BSA, pH 7.8 @ 25°C], incubated for the specified length of time at 16°C, and terminated by heat denaturation of the ligase.

3.2.6 Amplification of indexed bacterial genomic DNA fragments by Polymerase Chain Reaction (PCR) : standard conditions

Amplification of indexed bacterial genomic DNA fragments with *Taq* DNA polymerase was performed using PCR reagents and conditions previously described for the amplification of indexed pUC19 fragments. Amplification of IGP target fragments with *PfuTurbo*™ DNA polymerase was performed using reagents and conditions previously described for the amplification of indexed pUC19 sequencing tempts.

3.2.7 Agarose gel electrophoresis of amplified indexed bacterial DNA fragments on large 64-mix gels

Analysis of indexed bacterial genomic DNA fragments by agarose gel electrophoresis was performed in a manner similar to that outlined in **Chapter II**. Photodocumentation on Polaroid film or by digital camera was also performed as previously described.

3.2.8 Software development and database construction

The C++ program EcoliDB v1.0 (developed by Randy Nonay and Chris Dambrowitz) was developed to facilitate manipulation of bacterial genome sequence data in preparation for DNA indexing analysis. The complete *E. coli* genome sequence (*E. coli* subsp. K-12 strain MG1655 version M52, 4 639 221 bp, GenBank Accession #NC000913) was downloaded as an ASCII text file [155]. EcoliDB opened the file and searched the sequence for *Fok*I recognition site motifs in both the 5' and 3' direction. For each recognition site identified in the sequence, the program generated a set of data entries providing a complete indexing-based description of each predicted *Fok*I restriction fragment in an *E. coli* DNA digest. The data entries generated by the program were saved to a comma-delineated text file that could be manipulated by Microsoft Excel 98 (Microsoft Corporation, Redmond WA) to create a versatile searchable database of indexable fragments in *Fok*I-digested bacterial genomic DNA. This database was employed in the design of bacterial indexing investigations and in the analysis of experimental results.

3.3 RESULTS AND DISCUSSION

3.3.1 Distribution of *Fok*I restriction sites across the *E. coli* genome

The *E. coli* genome is composed of 4 639 221 bp with a G/C content of 50.8% [151]. Statistical expectations for the frequency of occurrence of a particular 5-bp sequence in a genome of this size and %GC content suggest that digestion of *E. coli* chromosomal DNA with *Fok*I endonuclease would yield roughly 9060 restriction fragments averaging 512 bp in length. The complete *E. coli* genome sequence was searched for *Fok*I recognition motifs using Sequencher v3.0 sequence data analysis software. Search results revealed the actual number of *Fok*I restriction sites in the genome to be 10 448, indicating the average size of *Fok*I fragments generated from *E. coli* chromosomal DNA to be roughly 450 bp. These results suggest that DNA indexing approaches based on *Fok*I restriction fragments provide excellent coverage of the *E. coli* genome for indexed profiling purposes. However, the possibility of significant regional bias in *Fok*I fragment coverage of the genome was not eliminated by these preliminary data. Further analysis of the genome sequence indicated that the largest *Fok*I fragment generated from chromosomal DNA (i.e. the largest genomic region lacking intervening *Fok*I recognition motifs) was expected to be 4.5 kb. Taken together, these data demonstrated excellent coverage of the *E. coli* genome by *Fok*I restriction fragments, without any significant regional bias. The uniform accessibility to all regions of the genome afforded by *Fok*I indicated that targeting of restriction fragment population subsets by a series of indexer combinations would allow indexing-based fingerprinting approaches employing this enzyme to objectively sample a cross-section of the entire bacterial genome at high resolution. *Fok*I endonuclease is therefore an appropriate choice of restriction enzyme for profiling of bacterial genomes by DNA indexing.

3.3.2 Amplification of indexed restriction fragments from *Fok*I-digested *E. coli* chromosomal DNA using pairs of phosphorylated indexers

Suitable ligation and PCR reaction conditions needed to be identified that would permit successful amplification of targeted restriction fragments from a DNA digest as complex as an entire bacterial genome. In order to establish preliminary benchmarks against which subsequent improvements in indexing reaction conditions could be evaluated, an initial foray into *Fok*I-digested *E. coli* chromosomal DNA was performed with pairs of phosphorylated Bam indexers. The Bam common primer was designed with the intention that it could be used to amplify properly-indexed fragments from digested *E. coli* DNA, without generating PCR products directly from *E. coli* DNA itself. The Bam primer sequence (and, therefore, its complementary sequence) is not present in *E. coli* DNA, and predictive modeling indicated that under appropriately-stringent PCR conditions no priming of spurious products directly from chromosomal DNA sequence was anticipated. Standard PCR reactions containing 100 ng of *Fok*I-digested *E. coli* DNA and 40 pmol of Bam primer did not generate amplification products (data not shown), confirming this prediction.

The *E. coli* genomic sequence was searched for *Fok*I fragments targeted by the cohesive end sequences of pairwise combinations of Bam P-indexers employed in development of the pUC19 model system. The availability of the *E. coli* genomic sequence provided a means to evaluate the accuracy of indexing in the context of a complex system. Fragments anticipated on the basis of the sequence data could be targeted and their amplification (and that of other, "unanticipated" fragments) could be confirmed or denied as a testable hypothesis. Searching the genome sequence for *Fok*I restriction sites generating a particular cohesive end sequence was performed using Sequencher v3.0 sequence data analysis software, followed by manual searching for adjacent *Fok*I end sequences and manual cataloging of search results.

From the manual mining of the sequence data for several fragment classes, two pairs of Bam P-indexers were selected which targeted specific fragments of known size and cohesive end sequence (FIGURE 3.1). More than a single pair of indexers were experimentally tested in order to increase the likelihood that at least amplified indexed

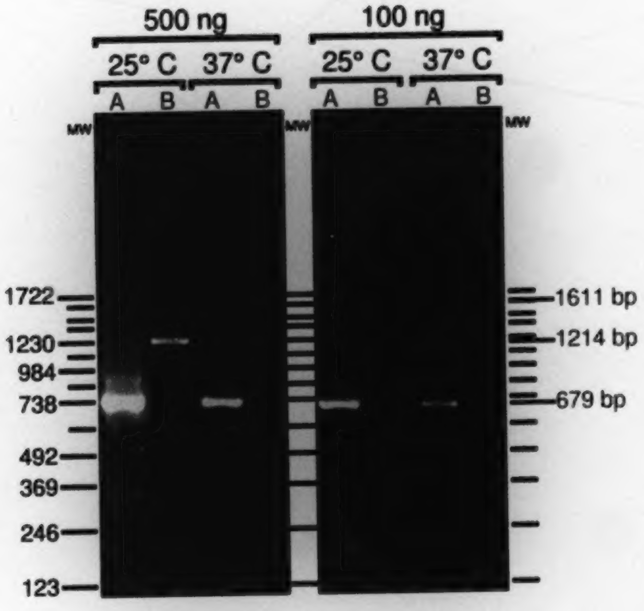


FIGURE 3.1: Amplification of indexed restriction fragments from *FokI*-digested *E. coli* chromosomal DNA.
Two pairs of Bam P-indexers were selected targeting specific fragments of known size and cohesive end sequence in a *FokI* digest of *E. coli* chromosomal DNA. Ligation of the chromosomal digest with P-GCGAxBam and P-TGTCxBam (indexer combination A, identified in red) was expected to generate two amplifiable fragments, 679 bp and 1611 bp in length. A single 1214-bp PCR product was anticipated for indexing reactions containing P-TCTTxBam and P-TCGCxBam (indexer combination B, noted in blue). Ligations employing 40 U T4 DNA ligase and containing either 500 ng or 100 ng *FokI*-digested *E. coli* DNA were assembled in 20-µl reaction volumes. Each indexer combination (50 fmol/indexer/ligation) was ligated to the two different amounts of chromosomal DNA at two different temperatures, 25°C and 37°C. Following 60 min of ligation at the appropriate temperature, the reactions were terminated, and aliquots added to standard indexing amplification reactions. Thirty cycles of standard-condition PCR was performed with *Taq* DNA polymerase using 40 pmol of Bam primer. Amplification reactions were analyzed by agarose gel electrophoresis. Columns of black bars marked "MW" represent 123-bp ladder molecular weight markers.

fragments would be generated. However the tedious manual searching approach limited the number of fragment classes which could be efficiently evaluated. Ligation of chromosomal digest with P-GCGAxBam and P-TGTCxBam was expected to generate two amplifiable fragments, 679 bp and 1611 bp in length. A single 1214-bp PCR product was anticipated for indexing reactions containing P-TCTTxBam and P-TCGCxBam. (Subsequent analysis of *E. coli* genome sequence data using DNA indexing database software showed that, fortuitously, none of the four P-indexers used in this experiment were specific for the generation of repeated-end fragments.)

Relatively high concentrations of *Fok*I-digested *E. coli* DNA (500 ng/ligation or 100 ng/ligation) were used in this experiment, in order to increase the probability that PCR products were obtained and establish basic benchmark conditions for DNA indexing from bacterial DNA. The 100-ng ligations were included to rapidly evaluate whether lower amounts of digest DNA might be used, reducing the rate at which materials were consumed. Each indexer combination (50 fmol/indexer/ligation) was ligated to the two different amounts of chromosomal DNA at two different temperatures, 25°C and 37°C. Ligation at 25°C was performed in each case in order to provide conditions that favoured the annealing of four-nt cohesive ends, thus increasing the probability that PCR products might be generated in this initial bacterial indexing attempt. Following 60 min of ligation with T4 DNA ligase, the reactions were terminated, and aliquots added to standard indexing amplification reactions. Thirty cycles of standard-condition PCR was performed with *Taq* DNA polymerase using 40 pmol of Bam primer.

As manual sequence data mining had predicted, indexing of *Fok*I-digested *E. coli* DNA with P-GCGAxBam and P-TGTCxBam produced two amplicons, 679 bp and 1611 bp respectively. A single 1214-bp indexed product was amplified from indexing ligations containing P-TCTTxBam and P-TCGCxBam. Ligation of indexers to 100 ng of chromosomal digest permitted amplification of indexed products in all cases, indicating the potential that lower amounts of target DNA might be sufficient. An incubation temperature of 37°C was not prohibitive for ligation of indexers to target fragments, and reduced the level of "background" amplification visible by agarose gel analysis. Thus for a range of reaction conditions, the targeted ligation of indexers to specific *Fok*I fragments in a complex bacterial genomic digest allowed successful amplification of the indexed target fragments.

3.3.3 Coverage of the *E. coli* genome with *Fok*I restriction sites generating the cohesive end sequences CGCG and GCGC

In contrast to indexing-based genomic mapping and sequencing efforts, molecular subtyping studies for species and strain differentiation do not require the complete description of a bacterial chromosome by DNA indexing. Only a small subset of the total number of restriction fragments in a chromosomal digest are required in order to generate strain-specific patterns of product amplification. Rather than a complete survey of every indexable fragment across each fragment class, a set of indexing ligations using a single P-indexer against a selection of NoP indexers provides sufficient discriminatory power to characterize bacterial strains. For initial demonstrations of bacterial genomic profiling by DNA indexing, the palindromic cohesive end sequences CGCG and GCGC were typically selected as targets for ligation with P-indexers (P-CGCGxBam and P-GCGCxBam, respectively). These cohesive end sequences both exhibited several characteristics that made them convenient choices for P-indexer targeting in fingerprinting experiments.

As both sequences are palindromic, *a Fok*I cleavage event that generates either CGCG or GCGC as a cohesive end on one end of a restriction fragment necessarily generates the identical sequence (again, CGCG or GCGC, respectively) on the complementary end of the adjacent *Fok*I fragment. Assaying a *Fok*I digest for a CGCG "hit" from a particular restriction site will produce two "hits" from the same site with the same P-indexer sequence (but presumably in different NoP-indexer ligation reactions) corresponding to adjacent fragments. This property makes the targeting of palindromic cohesive end sequences useful in indexing-based fingerprinting studies by effectively doubling the information content regarding a particular genomic region accessed in studies using one of the appropriate indexers.

The high %GC content of both typically-targeted cohesive end sequences offered increased probability that annealing between the indexer cohesive ends and those of the targeted end of the fragment would be sufficiently stable to favour P-indexer ligation. (Even in instances in which the GC-rich cohesive end sequence of the P-indexer was sufficiently stable as to favour mismatches, ligation fidelity is somewhat less critical in a fingerprinting study as compared to a mapping project, for example. The ability to generate highly-reproducible amplification product patterns is of greater importance in this context. However, while a reduced emphasis on ligation fidelity might be acceptable during actual application of indexing-based molecular subtyping, for the purposes of this demonstration of bacterial profiling, ligation fidelity was considered an essential criterion with which to evaluate indexing performance.)

Finally, it was determined from manual cataloging of Sequencher searches of the *E. coli* sequence data that no repeated-end fragments in *Fok*I-digested *E. coli* DNA would be generated using indexers targeting CGCG or GCGC cohesive end sequences. This prediction was again obtained by searching the genome sequence for *Fok*I restriction sites generating the cohesive end sequences CGCG or GCGC using sequence data analysis software, followed by manual cataloging of the cohesive end sequences of each of the two *Fok*I fragments generated at each targeted restriction site. These data were used to build a partial profiling map of the *E. coli* chromosome illustrating the predicted coverage of the genome with *Fok*I restriction sites generating CGCG and GCGC cohesive end sequences (FIGURE 3.2). The map demonstrates adequate coverage and "sampling" of genomic regions by either of these two cohesive end sequences. Similar levels of uniform genome coverage were identified for other cohesive end sequences (data not shown). Generally, then, the cohesive end sequences generated by *Fok*I digestion in *E. coli* chromosomal DNA are relatively uniformly distributed along the chromosome, providing an appropriate basis for demonstration of molecular fingerprinting of the *E. coli* genome by *Fok*I endonuclease.

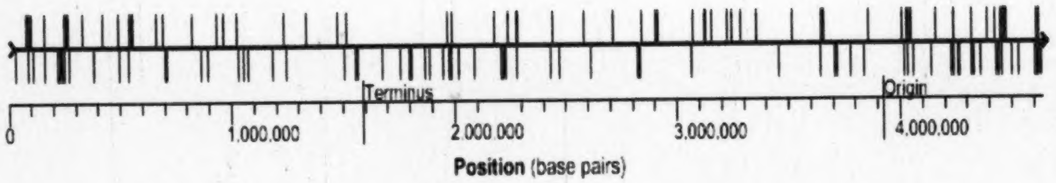


FIGURE 3.2: Coverage of the *E. coli* genome with *Fok*I restriction sites generating the cohesive end sequences CGCG and GCGC.
A *Fok*I cleavage event generating either CGCG or GCGC as a cohesive end on one end of a restriction fragment generates the identical sequence on the complementary end of the adjacent *Fok*I fragment. Searching the genome sequence for *Fok*I restriction sites generating the cohesive end sequences CGCG or GCGC using Sequencher v3.0 sequence data analysis software, followed by manual cataloging of the cohesive end sequences of each of the two *Fok*I fragments generated at each targeted restriction site, permitted the assembly of a partial profiling map of the *E. coli* chromosome illustrating the predicted coverage of the genome with *Fok*I restriction sites generating CGCG and GCGC cohesive end sequences. *Fok*I restriction sites generating CGCG cohesive ends are identified in blue, while *Fok*I sites producing GCGC cohesive ends are identified in red. The map demonstrates adequate coverage and "sampling" of genomic regions by either of these two cohesive end sequences alone or in conjuction with one another.

3.3.4 Evaluation of ligation and amplification reaction conditions for indexing-based profiling of bacterial genomes

3.3.4.1 Rationale for the use of pooled NoP indexer mixes in ligation for bacterial fingerprinting

Searches of the *E. coli* genomic DNA sequence had revealed there to be 10 448 *Fok*I restriction sites along the chromosome. Indexing systems based on two indexers with 4-nt informative end sequences permit the characterization of 32 896 fragment classes. Therefore, at least two-thirds of the fragment classes available in such an indexing system, and perhaps as many as 25 000 fragment classes, will be entirely unrepresented in a *Fok*I digest of *E. coli* DNA. In other words, there is less than a 1-in-3 chance that any given pair of 4-nt indexers will find a legitimate target in an indexing ligation, for bacterial genomes of comparable complexity. An approach to genomic profiling that is based on such an indexing system would clearly be an inefficient use of the informative capacity of DNA indexing.

The use of pools of NoP indexer sequences in ligations of bacterial genomic digests offers a more appropriate and efficient method of indexing for profiling applications. The pooled NoP indexers, in combination with a single P-indexer, effectively assay for a set of fragment classes simultaneously in a single indexing reaction. Over the collection of NoP indexer sets, all 256 four-base indexing end sequences are represented, thus allowing all fragment classes to be assayed in one-fourth the number of indexing reactions. For this demonstration of indexing-based bacterial genomic profiling, 64 pools of 4 BamCC NoP indexers each were assembled. Each pool contained NoP indexers presenting the same trinucleotide sequence in the first three base positions of their cohesive ends, and then either A, C, G, or T in the fourth position. In other words, NoP indexer mixes were of the format XXXN, where XXX was a trinucleotide shared by all indexers in the pool, and where N was either A, C, G or T. For example, indexer mix 3 contained the indexers OH-AACAxBamCC, OH-AACCxBamCC, OH-AACGxBamCC and OH-AACTxBamCC. The use of NoP indexer pools reduced the number of indexing reactions required for full description of a genomic profile targeted by a particular P-indexer sequence. In this way the time and resources expended in obtaining an informative and discriminatory genomic profile were significantly reduced. An additional benefit of the use of 64 NoP indexer mixes (or *64-mixes*) was the reduced number of oligonucleotide syntheses required to assemble a functionally-complete set of indexers capable of targeting all 256 *Fok*I-type cohesive sequences.

Use of pooled sets of indexers has both theoretical and functional consequences for indexing systems. Although the 64-mixes retain the ability to target each of the 256 4-nt cohesive end sequences, the ability to discriminate targeted end sequences is reduced to three of the four base positions in each cohesive end. This reduces the effective number of fragment classes from the 32 896 classes distinguishable with 256 P-indexers and 256 NoP indexers to 8224 "64-mix" fragment classes defined by 256 individual P-indexers and 64 pools of 4 NoP indexers each. For any one "known" P-indexer sequence, the number of "unknown" fragment end-classes (targeted by NoP indexers) which can be distinguished drops from 256 to 64. This effectively quadruples the average number of "hits", or indexed and amplified fragments, in any one lane. While for some applications this would be a waste of indexing's informative capacity and would complicate interpretation of indexing experimental results, use of 64-mixes is appropriate for bacterial molecular subtyping applications, as two-thirds of indexing reactions using individual indexers would otherwise be expected to be uninformative.

Functional consequences of the use of 64-mixes are related to the increase in ligation complexity produced by the presence of four times the amount of NoP indexer in each ligation reaction. In order to attain a similar level of ligation events per indexer sequence in a reaction with five indexer sequences (one P-indexer and four NoP indexers) as in a reaction with only two indexer sequences, each NoP indexer in the mixed ligation must be present at roughly the same concentration as the single NoP indexer in the 2-indexer ligation. As a result, the total NoP indexer concentration in the mixed ligation is four times that of the NoP indexer concentration in the 2-indexer ligation. The increase in ligation complexity due to the presence of 64-mixes in a reaction results in substantially more forcing conditions for ligation. A somewhat higher frequency of misligation events may occur for ligations performed in the presence of 64-mixes than would be anticipated from ligations employing an individual NoP indexer. While reduced misligation rates are critical for many indexing applications, in the context of genomic profiling moderate levels of misligation may be tolerated, should they occur. As the goal of bacterial profiling experiments is essentially to generate a strain-specific pattern of amplification products that is highly reproducible and is differentiable from patterns generated by the same technique from the DNA of other bacterial strains or species, the presence of small numbers of reproducible misligated fragments is not disastrous, and may in fact offer a moderate level of additional discriminatory power. However, for this demonstration of genomic profiling via indexing using the *E. coli* genome as a test case, the capacity to evaluate and optimize indexing strategies and reaction conditions relied on the ability to predict and identify particular amplification products. Therefore efforts were made to minimize the frequency of misligation events due to ligation complexity.

3.3.4.2 Evaluation of ligation conditions for bacterial profiling 1: indexer concentration, DNA concentration, and ligase concentration; comparison of T4 DNA ligase and Taq *DNA ligase*

Ligation conditions appropriate for bacterial genomic profiling were anticipated to be different than those employed for indexing of the pUC19 model system. The contribution to ligation complexity due to the use of 64-mixes was dwarfed by the complexity of the target DNA - a mixture of *Fok*I restriction fragments representing the entire *E. coli* genome.

The *E. coli* chromosome is over 1700 times the size of pUC19 and contains over 2000 times the number of *Fok*I fragments contained by the plasmid used as a DNA indexing model system. In other words, the ligation conditions evaluated during investigations of the model system were optimized for a DNA digest presenting 0.06%

the complexity of *Fok*I-digested bacterial genomic DNA. Additionally, some percentage of *Fok*I restriction sites in the *E. coli* genome were anticipated to generate fragments too small or, rarely, too large, to be effectively indexed, amplified in a competitive-PCR environment and analyzed by gel electrophoresis (see **Section 3.3.6.2**). These non-indexable fragments would nonetheless contribute to ligation complexity. Finally, the pattern of amplified products generated from PCR reactions each containing indexed fragments belonging to four related fragment classes provides an increase in complexity, not of ligation but of interpretation of data.

In contrast to the simple case of pUC19, the successful production of indexed double-fragment amplicons is unlikely in a Type IIS restriction digest as complex as a bacterial genome. Religation (or *de novo* ligation) of two *Fok*I fragments with complementary ends, followed by ligation of a correctly-matched P-indexer to one exterior end and a correctly-matched NoP indexer or P-indexer to the other exterior end of the double-fragment, is highly unlikely. Given the large number of different fragments present in an indexing ligation of *E. coli* DNA, the probability is extremely low that a sufficient number of copies of a particular double-fragment with two indexed ends will be generated to allow amplification from a competitive-PCR environment. In this case, at least, the complexity of a bacterial genomic digest offers a reduced probability of generating an indexing artifact, relative to that of the simple pUC19 model system.

Taken together, the numerous factors contributing to complexity in indexing reactions featuring NoP indexer mixes and bacterial genomic digests increased the probability that reaction conditions, if improperly selected, could be strongly forcing for ligation and would generate an undesirable level of misligation events. In order to effectively evaluate the characteristics of the indexing system employed in this demonstration of bacterial genomic profiling, experimental results needed to be correlated with sequence-analysis-based predictions. Consequently, ligation conditions which minimized the effect of ligation complexity, and thus simplified interpretation of experimental results, were desirable.

In order to evaluate a range of ligation conditions for their applicability to indexed genomic profiling, a set of fragments predicted from manual cataloging of genome sequence search results were targeted by indexing ligations and the experimental results evaluated by comparison with those predictions (FIGURE 3.3). As previously mentioned, ligations of *Fok*I-digested *E. coli* DNA containing P-CGCGxBam were not expected to generate indexed repeated-end fragments. It was now established by the search methodology outlined above that the P-CGCGxBam/OH-GCGNxBamCC indexer combination was predicted to target three fragments: 1328 bp, 392 bp and 231 bp in length, respectively.

This indexer combination was used for each set of ligation conditions evaluated. The evaluated ligation parameters included NoP indexer mix concentration, *E. coli* DNA digest concentration, and ligase concentration. These parameters were evaluated for ligation both with T4 DNA ligase and for the prokaryotic enzyme *Taq* DNA ligase.

Ligations were performed in a total volume of 20 µl using the appropriate concentrations of OH-GCGNxBamCC NoP indexer mix, *Fok*I-digested *E. coli* DNA, and ligase. Each reaction contained 50 fmol of the P-indexer P-CGCGxBam. Ligations containing T4 DNA ligase were assembled using commercial T4 DNA Ligase Buffer, incubated for 1 h at 37°C, and terminated by heat-denaturation of the ligase prior to the aliquoting of 2 µl into a PCR reaction. Ligations containing *Taq* DNA ligase were assembled using commercial *Taq* DNA Ligase Buffer, incubated for 1 h at 42°C, and terminated by cooling the reaction to 0°C, at which point a 2-µl aliquot was removed for amplification. PCR was performed as previously described and the amplification products analyzed by agarose gel electrophoresis.

To ensure that amplification products were generated in this initial demonstration featuring indexed *E. coli Fok*I fragments targeted by a NoP indexer mix and a P-indexer, high concentrations of T4 DNA ligase (400 U/reaction) were added to certain ligations. Although most users of T4 DNA ligase routinely employ such enormous amounts of enzyme (for ligating insert DNA into a vector, for example), for indexing purposes this amount of ligase is excessive, even in a reaction containing as

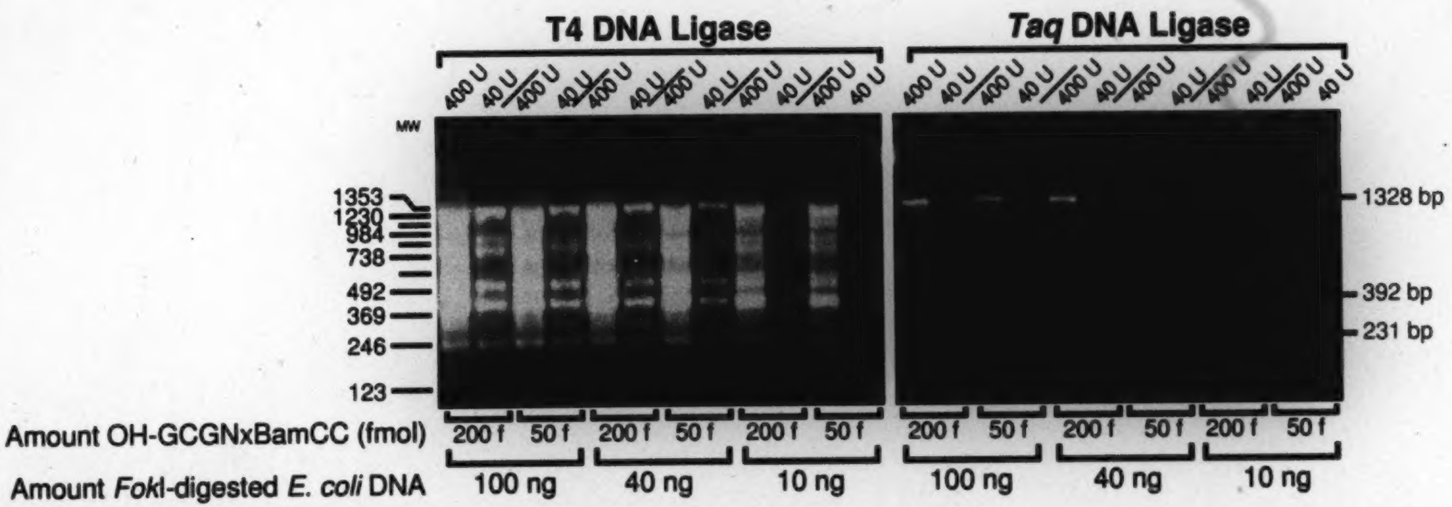

T4 DNA Ligase
Taq DNA Ligase
400 U
40 U
MW
1353
1230
984
738
492
369
246
123
1328 bp
392 bp
231 bp
Amount OH-GCGNxBamCC (fmol)
200 f
50 f
Amount FokI-digested E. coli DNA
100 ng
40 ng
10 ng

FIGURE 3.3: Evaluation of ligation conditions for bacterial profiling 1: Indexer concentration, DNA concentration, and ligase concentration; comparison of T4 DNA ligase and *Taq* DNA ligase.
The P-CGCGxBam/OH-GCGNxBamCC indexer combination was predicted from manual cataloging of genome sequence search results to target a set of three fragments 1328 bp, 392 bp and 231 bp in length. This indexer combination was used for each set of ligation conditions evaluated.
Ligations were performed in a total volume of 20 µl using the appropriate concentrations of OH-GCGNxBamCC NoP indexer mix, *Fok*I-digested *E. coli* DNA, and ligase. Each reaction contained 50 fmol of the P-indexer P-CGCGxBam. Ligations containing T4 DNA ligase were assembled using commercial T4 DNA Ligase Buffer, incubated for 1 h at 37°C, and terminated by heat-denaturation of the ligase prior to the aliquoting of 2 µl into a PCR reaction. Ligations containing *Taq* DNA ligase were assembled using commercial *Taq* DNA Ligase Buffer, incubated for 1 h at 42°C, and terminated by cooling the reaction to 0°C, at which point a 2-µl aliquot was removed for amplification. PCR was performed as previously described and the amplification products analyzed by agarose gel electrophoresis. The columns of black bars marked "MW" represents 123-bp ladder molecular weight markers.

many different potential ligase substrates as a restriction digest of whole genomic DNA. In this experiment, ligation reactions employing 40 U of T4 DNA ligase exhibited significantly lower levels of "background" amplification products following PCR, while those employing 400 U gave rise to enough background amplification to completely obscure any legitimate pattern of indexed product amplification in some cases.

With so many more fragments in a ligation of *Fok*I-digested bacterial DNA than in a ligation of pUC19 fragments, and with the dramatic increase in the number of fragments with a sequence targeted by a NoP indexer of the 64-mix and no sequence complementary to the featured P-indexer at the other, it might have been expected that a greater amount of each specific NoP indexer would be required to enable amplification of an indexed fragment than for a specific NoP indexer in a pUC19 ligation. For this reason, certain ligations were performed using a higher concentration of each indexer (200 fmol/indexer/reaction, for a total of 800 fmol/reaction total NoP indexer). Other ligations included only 50 fmol/indexer/reaction (200 fmol/reaction total NoP indexer). Contrary to what might have been expected, 50 fmol of each NoP indexer per ligation is sufficient to allow the amplification of indexed fragments from an *E. coli Fok*I digest.

Three different concentrations of chromosomal DNA digest were evaluated in this experiment: 100 ng/reaction, 40 ng/reaction, and 10 ng/reaction. Despite the three-order-of-magnitude increase in complexity between pUC19 and *E. coli* indexing systems, only a tenfold increase in total digest concentration (i.e. 10 ng/ligation) is required in order to amplify indexed *Fok*I fragments. Effective amplification of each of the three predicted target fragments is evident from ligations containing as little as 10 ng *Fok*I-digested *E. coli* DNA, 50 fmol/ligation of specific NoP indexer (200 fmol total), and 40 U of T4 DNA ligase. These ligation conditions, surprisingly similar to those used in the development of the pUC19 model system, generate little background amplification due to misligation. (The presence of a specific, reproducible amplification product roughly 500 bp in length in this lane is likely a legitimately-indexed "unanticipated" fragment, as discussed in **Section 3.3.6.2.**) In addition, these ligation conditions permit efficient use of indexing resources, reducing the cost of performing IGP investigations in a clinical or service-lab setting.

Taq DNA ligase, more accurately known as *Thermus thermophilus* DNA ligase [156], is a thermostable eubacterial NAD$^+$-dependent ligase capable of very high fidelity under certain ligation conditions [39]. The enzyme's nominal optimum temperature for ligation efficiency is 45°C [39], a temperature at which 4-nt cohesive ends are unable to anneal effectively. The potential for higher ligation fidelity in IGP reactions employing *Taq* DNA ligase was investigated.

To evaluate whether use of *Taq* DNA ligase was appropriate for IGP applications, a set of ligation reactions covering the range of DNA concentrations, NoP indexer mix concentrations and ligase concentrations evaluated by T4 DNA ligase was prepared. Ligations using *Taq* DNA ligase were performed at 42°C, in order to provide a balance between the enzyme's optimal activity temperature and the ability of four-base cohesive ends to anneal to one another with even minimal stability. The incubation time suggested for *Taq* DNA ligase by the supplier were much longer (12 hours or more) than that required for efficient ligation with T4 DNA ligase. For envisioned applications of IGP studies, such as in clinical or service-lab settings where experimental throughput is a consideration, the time required per IGP study is an important factor. For this reason, ligation times for *Taq* DNA ligase were kept to the same 1-hr incubation period as used for T4-ligase-based reactions.

To a certain extent, *Taq* DNA ligase's reputed capacity for high-fidelity ligations was borne out by the results of this investigation. The level of background amplification products generated by PCR from *Taq*-ligase-based indexing of *Fok*I-digested *E. coli* fragments was much lower in reactions containing large amounts (400 U) of enzyme than that generated from ligations with similar amounts of T4 DNA ligase (using the same functionally-defined Units for both enzymes). However, at the low DNA and indexer concentrations (relative to typical non-indexing molecular biology applications) efficiently ligated by the phage enzyme, ten times as much *Taq* DNA ligase (400 U) was required to achieve the same amount of amplifiable indexed product. Even at such high concentrations, the pattern of amplification products generated by *Taq* DNA ligase was very similar to that generated by T4 DNA ligase (suggesting that at, under indexing ligation conditions, the fidelity of T4 DNA ligase is as good or better than that of the eubacterial enzyme). In fact, *Taq* DNA ligase was not capable of efficiently indexing the smallest (231-bp) predicted target fragment at any except the highest concentrations of ligase, target DNA and NoP indexer.

From the perspective of presenting IGP studies as an attractive method for bacterial fingerprinting for clinical and service labs, a simple cost analysis further discriminated against *Taq* DNA ligase as a candidate ligase for IGP applications. T4 DNA ligase costs less than 1/4¢ per unit of enzyme, or10¢/ligation (for a 40 U ligation). *Taq* DNA ligase costs 3.25¢ per unit of enzyme, or $1.30 per 40-U ligation. In order to achieve the same ligation efficiency as 40 U of T4 DNA ligase in 1 hour of ligation, 400 U of *Taq* DNA ligase would have been required, bringing the cost of each ligation to $13. While it was decided on the basis of these results that *Taq* DNA ligase was not an appropriate choice of enzyme for IGP applications, these data demonstrated the potential of *Taq* DNA ligase for use as an indexing ligase for certain applications (as discussed in **Section 4.3.12**).

3.3.4.3 Evaluation of ligation conditions for bacterial profiling 2: ligation temperature, ligation time and ligase concentration; comparison of T4 DNA ligase and E. coli *DNA ligase*

In order to further refine the reaction conditions established for IGP ligations employing T4 DNA ligase, reaction parameters including incubation temperature, incubation time and ligase concentration were evaluated (FIGURE 3.4). Evaluation of *E. coli* DNA ligase for use as a potential indexing enzyme for IGP applications was also performed. A pair of ligations was performed with T4 DNA ligase at 37°C for 1 hr using the standard ligase, target DNA and indexer concentrations established as benchmarks in the previous experiment. A second pair of similar ligations was incubated for 12 h for comparison both with the benchmark conditions and with 12 h incubations using *E. coli* DNA ligase. (Despite the relative impracticality of a IGP protocol requiring 12-h ligations, efforts were made to discover if *E. coli* DNA ligase

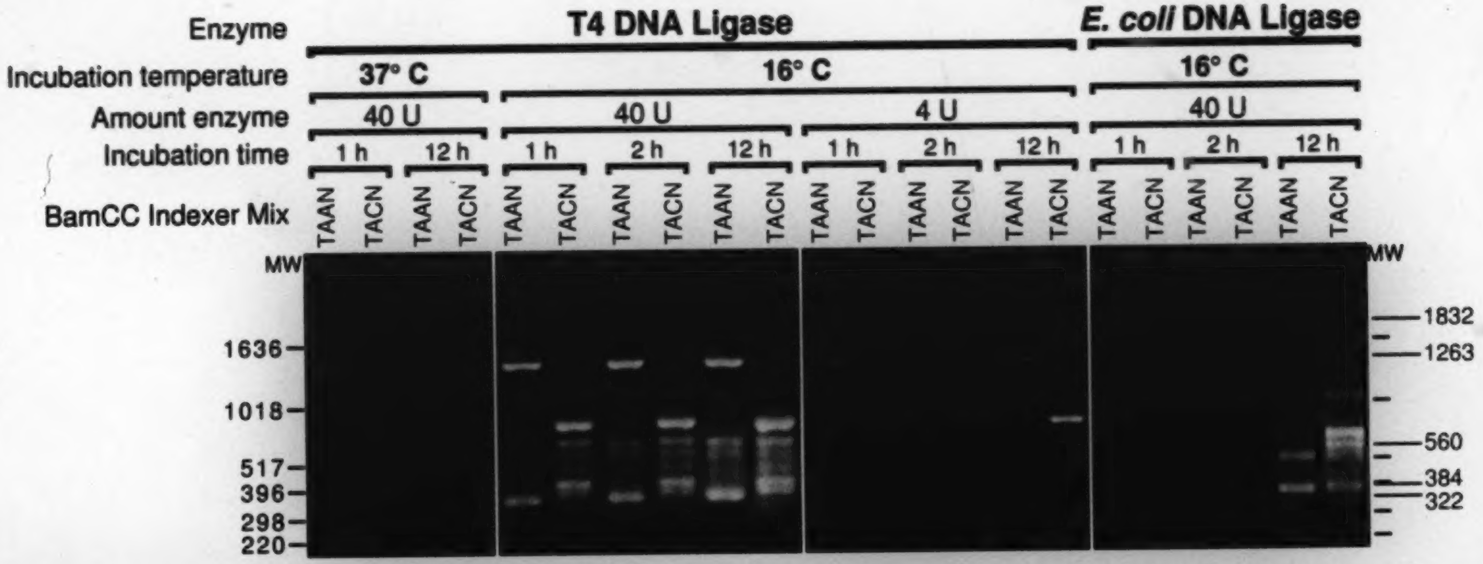

Enzyme
Incubation temperature
Amount enzyme
Incubation time
BamCC Indexer Mix
T4 DNA Ligase
E. coli DNA Ligase
37° C
16° C
16° C
40 U
40 U
4 U
40 U
1 h
12 h
1 h
2 h
12 h
1 h
2 h
12 h
1 h
2 h
12 h
TAAN
TACN
MW
MW
1636
1018
517
396
298
220
1832
1263
560
384
322

FIGURE 3.4: Evaluation of ligation conditions for bacterial profiling 2: ligation temperature, ligation time and ligase concentration; comparison of T4 DNA ligase and *E. coli* DNA ligase.
Two sets of fragments were targeted by indexing ligations for evaluation of ligation temperature, ligation time and ligase concentration, and for comparison of T4 DNA ligase and *E. coli* DNA ligase. Manual cataloging of genome sequence search results determined that the P-GCGCxBam/OH-TAANxBamCC indexer combination (noted in red) was predicted to target three restriction fragments from a *Fok*I digest of the *E. coli* chromosome, generating indexed products of 1263 bp, 560 bp and 322 bp, respectively. The P-GCGCxBam/OH-TACNxBamCC indexer combination (noted in blue) was predicted to produce two amplicons of 1832 bp and 384 bp. Each indexer combination was employed for each set of ligation conditions evaluated with either T4 or *E. coli* DNA ligase. Ligations were performed in a total volume of 20 µl using the appropriate concentrations of the specified ligase in the appropriate reaction buffer. Each reaction contained 50 fmol of the P-indexer P-GCGCxBam and 50 fmol/ligation/indexer (200 fmol total indexer) of the appropriate NoP indexer mix. Ligations were incubated for the appropriate time at the appropriate temperature, and terminated by heat-denaturation of the ligase prior to the aliquoting of 2 µl into a PCR reaction. PCR was performed as previously described and the amplification products analyzed by agarose gel electrophoresis. Columns of black bars marked "MW" represent 1-kb ladder molecular weight markers (Gibco-BRL).

would demonstrate potential as an indexing enzyme under any conditions. In fact, in ligations incubated for more than 1 h at 37°C, the T4 DNA ligase is likely to be inactivated *in vitro* due to the elevated temperature.) Ligations with varying incubation times were performed at 16°C with 40 U of T4 DNA ligase for comparison with results of similar ligations employing *E. coli* DNA ligase, and to observe the effect of lowered incubation temperature on ligation fidelity for indexing of bacterial genomic DNA digests. Similar ligations employing 4 U of T4 DNA ligase were included to evaluate whether efficient ligation of reasonable fidelity, generating low background amplification, could be achieved with smaller amounts of enzyme at lower incubation temperatures.

Like *Taq* DNA ligase, *E. coli* DNA ligase is a member of the NAD^+-dependent eubacterial ligase family [34, 157, 158]. As such, it has the reputation of being capable of high-fidelity ligation under certain ligation conditions [159]. Unlike *Taq* DNA ligase, however, *E. coli* ligase is not thermostable and in fact exhibits optimal ligation efficiency at 16°C, a temperature at which 4-nt cohesive ends are generally capable of effective annealing to one another. The potential for higher ligation fidelity in IGP reactions employing *E. coli* DNA ligase was investigated.

Two sets of fragments were targeted by indexing ligations for evaluation of experimental results. As previously discussed, the P-indexer P-GCGCxBam was not expected to generate indexed repeated-end fragments from ligations containing *Fok*I-digested *E. coli* DNA. Manual cataloging of genome sequence search results determined that the P-GCGCxBam/OH-TAANxBamCC indexer combination was predicted to target three FokI restriction fragments, generating indexed products of 1263 bp, 560 bp and 322 bp, respectively. The P-GCGCxBam/OH-TACNxBamCC indexer combination was predicted to produce two amplicons of 1832 bp and 384 bp. Each indexer combination was employed for each set of ligation conditions evaluated with either T4 DNA ligase or *E. coli* DNA ligase.

Ligations were performed in a total volume of 20 µl using the appropriate concentrations of ligase in the appropriate reaction buffer. Each reaction contained 50 fmol of the P-indexer P-GCGCxBam and 50 fmol/ligation/indexer (200 fmol total indexer) of the appropriate NoP indexer mix. Ligations were incubated for the appropriate time at the appropriate temperature, and terminated by heat-denaturation of the ligase prior to the aliquoting of 2 µl into a PCR reaction. PCR was performed as previously described and the amplification products analyzed by agarose gel electrophoresis.

High levels of misligation and background amplification were generated in ligations incubated at 16°C with 40 U of T4 DNA ligase. The use of 4 U of enzyme to compensate for the lower incubation temperature did not provide sufficient relief from misligation to be practical, and demonstrated insufficient ligation efficiency to permit amplification of several predicted target fragments. Generally, extended incubation times for T4-ligase-based ligations provided little or no increase in the amount of amplified product generated from the reaction. None of the reaction conditions evaluated demonstrated improved ligation characteristics relative to the benchmark conditions for IGP reactions established in the previous experiment.

E. coli DNA ligase was evaluated as a potential indexing enzyme for IGP applications largely because incubation at the enzyme's optimal activity temperature (16°C) would provide more suitable conditions for 4-nt cohesive end sequences to anneal to one another than those provided at *Taq* DNA ligase's incubation temperature of 42°C. It was hoped that predicted target fragments refractory to indexing with *Taq* DNA ligase would be indexed with *E. coli* DNA ligase, and that the reputed ligation fidelity of *E. coli* ligase would provide sufficient specificity to reduce the amount of background and misligation otherwise generated at this low an incubation temperature. This hypothesis was not borne out by the results of this experiment. As *E. coli* DNA ligase requires 12-h incubation periods in order to achieve an effective level of ligation, ample opportunity is provided for numerous misligation events to occur at such low temperature in such a complex mixture under such forcing conditions. *E. coli* DNA ligase was rejected as a potential indexing enzyme for IGP applications.

Even if *E. coli* DNA ligase had demonstrated promise as an indexing enzyme, the cost of its use in a clinical or service lab setting would be prohibitive. *E. coli* DNA ligase is more than 100x more expensive per unit than T4 DNA ligase. T4 DNA ligase costs less than 1/4¢ per unit enzyme, or10¢/ligation (for 40 U of enzyme). *E. coli* DNA ligase costs 25¢ per unit enzyme or $10.00 per 40 Unit ligation. Additionally, IGP studies employing *E. coli* DNA ligase would require over three times as long to perform, including ligation, PCR and amplification product analysis. This would cut sample throughput in half, even by two-thirds if the clinical lab IGP system was automated for 24-hour sample throughput.

3.3.5 Amplification of IGP ligation reactions

3.3.5.1 Background amplification patterns from ligations predicted to be devoid of legitimate indexing targets

The generation of spurious amplification products (or "background") by PCR from complex ligation mixtures such as those present IGP reactions was investigated. Through manual cataloging of Sequencher search results from the *E. coli* genomic sequence, several combinations of NoP indexer mixes (TAGN, TATN and TCAN) with the P-GCGCxBam P-indexer were identified that were not predicted to target any fragment in a *Fok*I digest of *E. coli* chromosomal DNA. Indexing reactions "targeting" these predicted vacant fragment classes were performed in order to observe the amplification products generated from such ligations.

As part of the same investigation, several NoP indexer mixes (GTGN, TCTN, TGTN, and TTGN) that were predicted to target *E. coli* restriction fragments when used in ligation in combination with P-GCGCxBam were also identified. In this case, neither the total number of fragments targeted per ligation, nor each of the targeted fragment lengths, was determined. Due to the tedious nature of the manual cataloging approach to fragment prediction from the published sequence, determination that a particular NoP indexer mix, in combination with P-GCGCxBam, would target at least one genomic restriction fragment was sufficient for inclusion of that NoP indexer mix in the "targeted fragment" portion of this experiment. [Once software had been designed that permitted the assembly of a database of predicted *Fok*I fragments directly from bacterial genomic sequence data, it became possible to easily predict the number of fragments targeted by each indexer combination, in addition to the lengths of each of those fragments (see **Section 3.3.6.1**).]

Ligations were performed using the benchmark reaction conditions determined for IGP reactions. Specifically, 20-µl reaction volumes using 50 fmol of P-GCGCxBam P-indexer, 50 fmol/indexer (200 fmol total) of the appropriate NoP indexer mix, 10 ng *Fok*I-digested *E. coli* DNA, and 40 U of T4 DNA ligase were assembled. Ligations were incubated for 1 h at 37°C, and terminated by heat-denaturation of the ligase prior to the aliquoting of 2 µl into a PCR reaction. PCR was performed using either *Taq* DNA polymerase or *PfuTurbo*™ DNA polymerase (see below) as previously described and the amplification products analyzed by agarose gel electrophoresis (FIGURE 3.5).

Generation of spurious "background" amplification products from ligations which did not target "legitimate" *Fok*I fragments was anticipated for the complex fragment mixtures found in IGP ligations. It had been demonstrated that the Bam primer sequence alone did not produce amplification products from unindexed *E. coli* DNA. Additionally, no repeated-end fragment bearing the cohesive end sequence GCGC were predicted from the genomic sequence. Therefore, any products amplified from ligations predicted to be devoid of correctly-targeted fragments were either misligation products or fragments generated by aberrant *Fok*I cleavage (due to proximity of adjacent recognition sites). With no legitimate, effectively-indexed fragment present in sufficient copy numbers competing for efficient amplification, low levels of these background amplicons become targets for the tremendous amplification power of PCR.

The pattern of background amplification products generated from ligations containing the three "predicted vacant" indexer combinations exhibited a number of amplification products common to all reactions, and several amplification products specific to particular NoP indexer mixes. The amplification products common to all three reactions are specific for all "vacant ligations" employing a particular P-indexer sequence. Different P-indexer sequences generated different conserved background amplification product patterns from a particular DNA source (data not shown). These

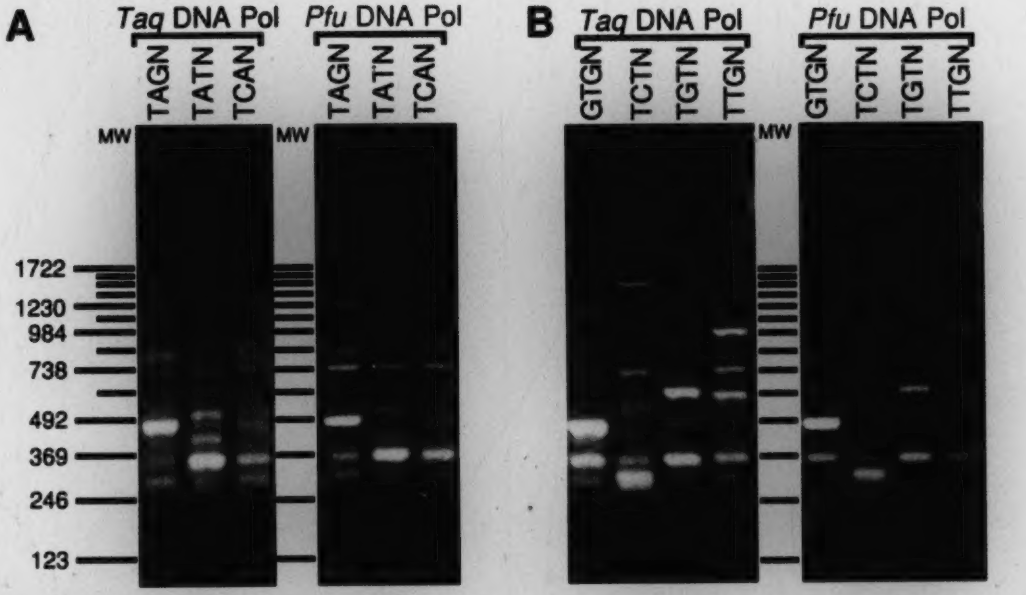


FIGURE 3.5: Background amplification patterns from ligations predicted to be devoid of legitimate indexing targets.
A) Manual cataloging of search results from the *E. coli* genomic sequence identified several combinations of NoP indexer mixes (TAGN, TATN and TCAN) with the P-GCGCxBam P-indexer that were predicted to target no fragment in *Fok*I-digested *E. coli* chromosomal DNA. Indexing reactions "targeting" these fragment classes were performed in order to observe the amplification products generated from such ligations.
B) Several NoP indexer mixes (GTGN, TCTN, TGTN, and TTGN) predicted to target at least one *E. coli* restriction fragment when used in combination with P-GCGCxBam were also identified by the same search methodology employed in part A).
Ligations for both part A) and B) were performed using the benchmark reaction conditions determined for IGP reactions. Specifically, 20-µl reaction volumes using 50 fmol of P-GCGCxBam P-indexer, 50 fmol/indexer (200 fmol total) of the appropriate NoP indexer mix, 10 ng *Fok*I-digested *E. coli* DNA, and 40 U of T4 DNA ligase were assembled. Ligations were incubated for 1 h at 37°C, and terminated by heat-denaturation of the ligase prior to the aliquoting of 2 µl into a PCR reaction. PCR was performed using either *Taq* DNA polymerase or *PfuTurbo*™ DNA polymerase as previously described and the amplification products analyzed by agarose gel electrophoresis. Columns of black bars marked "MW" represent 123-bp ladder molecular weight markers.

P-indexer-specific patterns were the result of the amplification of specific, reproducibly-indexable fragments carrying either a misligated P-indexer at both ends, or a correctly-ligated P-indexer at one end and a misligated P-indexer at the other. The reaction-specific bands were derived from misligations or unanticipated targets of the P-indexer and any of the four NoP indexers present in the 64-mix used in that reaction. Despite the fact that all of these bands are unanticipated from the sequence data and some are even misligations, they are still useful in a molecular subtyping context as they provide an additional level of discriminatory power to IGP studies.

Ligations generating legitimate indexed amplicons from predicted *Fok*I target fragments were amplified without substantial levels of background amplification products. In PCRs of these reactions, the targeted amplicons significantly outcompeted spurious products for amplification. As a result, neither P-indexer-specific or indexer-combination-specific background amplification patterns interfered with indexed target amplification, and background was not present at levels which could complicate interpretation of IGP results.

3.3.5.2 Comparison of amplification characteristics of Taq *DNA polymerase and* PfuTurbo™ *DNA polymerase for IGP applications*

Parallel PCR reactions containing identical amounts of each ligation of this experiment were assembled using *Taq* DNA polymerase and *PfuTurbo*™ DNA polymerase in order to compare the level of background amplification generated by each enzyme. In PCRs of ligations targeting legitimate indexable fragments, *PfuTurbo*™ DNA polymerase was observed to generate lower levels of background amplification while providing efficient amplification of legitimate amplicons. As a result, *PfuTurbo*™ DNA polymerase was selected as the enzyme of choice for IGP amplifications.

3.3.6 Preliminary evaluation of correlation between predicted and detected indexing target fragments for indexed genomic profiling approaches

3.3.6.1 Software development and database construction

Preliminary attempts to identify predicted target fragments for IGP protocol evaluation involved searching the genome sequence for *Fok*I restriction sites generating a particular cohesive end sequence using Sequencher v3.0 sequence data analysis software, followed by manual searching for adjacent *Fok*I end sequences and manual cataloging of search results. This approach to fragment prediction for IGP method evaluation was tedious, time-consuming and subject to human error. The C^{++} program EcoliDB v1.0 was developed to facilitate manipulation of bacterial genome sequence data for DNA indexing analysis.

EcoliDB searched a text file containing the entire genome sequence of *E. coli* MG1655 for *Fok*I recognition site motifs in both the 5' and 3' direction. For each recognition site identified in the sequence, the program generated a set of data entries describing the counted number of *Fok*I cutsites from the start of the genome sequence; the base position number of the start of the cutsite; the distance in bases from the previous cutsite to the current cutsite; and the sequence of the four-base cohesive end generated by cleavage of downstream DNA by *Fok*I from that recognition site. From these data, the program generated a second set of data entries corresponding to bacterial genomic fragments generated by *Fok*I restriction endonuclease: a fragment label corresponding to the number of fragments counted from the beginning of the sequence; the base position numbers corresponding to either end of the fragment; the length of the fragment in bases including the two four-base cohesive ends; and the sequences of the indexers required to target the cohesive end sequences at the left end and the right end of the fragment, respectively. The data entries generated by the program were saved as comma-delineated text in a second file labeled "cutsites.txt". This Excel-readable file was used to create a versatile searchable database of indexable fragments in *Fok*I-digested bacterial genomic DNA. This database was employed in the design of IGP investigations and in the analysis of experimental results.

3.3.6.2 *Use of EcoliDB v1.0 to identify predicted indexable* Fok*I restriction fragments in* E. coli *chromosomal DNA*

The database was searched to identify the percentage of *Fok*I fragments predicted to be generated from an *E. coli* chromosomal digest that were likely to be amplifiable by DNA indexing. Fragments with lengths falling at either extreme of the fragment size distribution for *Fok*I digests may not be amplifiable by standard indexing PCR conditions to product levels sufficient for visualization. In certain applications in which a complete description of a *Fok*I fragment population is necessary, PCR conditions and DNA polymerases may be selected to increase the amplification efficiency of these indexed products. In the context of genomic profiling, however, such measures are usually unnecessary, provided that the number of amplifiable indexed target fragments is sufficiently descriptive of the initial target fragment population to allow discrimination between amplification product patterns generated for different bacterial strains.

As had been previously suggested by manual cataloging of Sequencher search results, the database search confirmed that *Fok*I digestion of the *E. coli* genome was predicted to generate only a single fragment larger than 4 kb in length (4523 bp). An additional 15 predicted fragments between 3 and 4 kb were identified, and 128 *Fok*I fragments were identified in the database as being between 2 and 3 kb in length. Amplification of indexed fragments by *Taq* DNA polymerase under standard indexing PCR conditions had been observed experimentally to be difficult for fragment amplicons larger than 2.5 kb. This effect was particularly common in competitive amplification reactions containing more than a single indexed target fragment. As a result, in some cases these large fragments may not be effectively detected during standard indexing analysis of a genomic digest. Fifty-three *Fok*I fragments greater than 2.5 kb in length were predicted by the database to be present in *Fok*I digests of *E. coli* chromosomal DNA, representing 153 957 bp of genetic information and composing 3.3% of the *E. coli* genome.

As in the case of extremely large restriction fragments, extremely small *Fok*I fragments may be refractory to amplification by DNA indexing. The presence of two

*Fok*I recognition sequences within 30 bp of one another may have the consequence that only one or the other, but not both sites, will be cleaved by *Fok*I on a particular DNA molecule. This is a consequence of the DNA structural requirements for *Fok*I endonuclease to bind to its recognition motifs and cleave the DNA substrate [76], and is dependent on the orientation of the two adjacent motifs with respect to one another. Failure of *Fok*I to cleave at one site results in a restriction fragment that extends beyond the cleavage location predicted in the database from the genomic sequence to the next *Fok*I restriction site. As a result, the restriction fragment generated is of a different length and (probably) bears a different cohesive end sequence at its distal end than that predicted in the database. The "virtual fragment" predicted in the database is not present in an indexing reaction targeting its amplification, and the actual restriction fragment is targeted by a combination of indexers for which it is not predicted as an amplification product. (These "unanticipated" fragments, not predicted from the genome sequence, are nonetheless legitimate indexing targets. Although their amplification significantly reduces the level of correlation between target fragment predictions in the database and the number of products amplified by a particular combination of P-indexer and NoP indexer mixes, they do not detract from the utility of DNA indexing for molecular subtyping, as they provide an additional source of discriminatory power in IGP studies of various bacterial strains or species.) In addition, although fragments greater than 30 bp are expected to cleave in a manner predicted by the database, experimental observation has indicated that fragments less than 60 bp are not easily indexed, amplified and resolved on agarose gels. Fragments less than 60 bp in length were therefore not expected to be reliable targets for indexing by the IGP approach. A survey of the database revealed that *Fok*I-digested *E. coli* DNA was predicted to include 1272 fragments of 60 bp or less, representing 41 510 bp, or less than 1%, of the *E. coli* genome.

The EcoliDB database was used to predict the anticipated percentage of indexable restriction fragments represented in a genomic *Fok*I restriction digest. The 53 large fragments identified as potentially resistant to amplification by DNA indexing represent 153 957 bp of genetic information, or 3.3% of the *E. coli* genome. *Fok*I-digested *E. coli* DNA was predicted to contain 1272 fragments that are 60 bp or less, representing 0.9 % of the E. coli genome. Therefore, even if the genetic information in very large and very small fragments refractory to indexing were to be discounted, genomic profiling by DNA indexing was still expected to provide access to 9123 *Fok*I fragments representing over 95% of the *E. coli* genome. Such predictive analysis of the *E. coli* genome sequence data was feasible due to the utility of the EcoliDB software and database.

3.3.6.3 Preliminary experimental evaluation of correlation between predicted and detected indexing target fragments

A simple preliminary evaluation of the amplification of targeted *Fok*I restriction fragments predicted in EcoliDB was performed. This provided an opportunity to confirm the ability of the database to identify *Fok*I restriction fragments which could be targeted by a particular indexer combination and amplified. Using the database, eight NoP indexer mixes were identified that were predicted to target specific *Fok*I fragments in ligations using P-GCGCxBam as the P-indexer. The number and indexed length of the fragments targeted by each NoP indexer mix were summarized (TABLE 3.1). Database entries were correlated with the genomic region targeted by each predicted fragment (Protein-Coding Genes Feature Table, *E. coli* Genome Database, Entrez – Genome [160]). The ease of access provided by IGF approaches to underlying sources of genetic variation demonstrated the potential to design typing studies that specifically target chromosomal regions of interest in combination with a more generalized fingerprinting strategy.

Eight ligations, each containing P-GCGCxBam and the appropriate NoP indexer mix, were assembled. Ligation was performed using the standard IGP benchmark ligation conditions, followed by PCR and amplified product analysis by agarose gel electrophoresis (FIGURE 3.6). Each of the target fragments predicted for each of the eight reactions was accurately indexed and amplified to a level sufficient for visualization on an IGP gel. There was significant variation in the amount of each predicted amplicon produced, and in all cases the amplification reactions contained a

TABLE 3.1: Anticipated targeting of *E. coli* chromosomal *Fok*I fragments by selected indexer combinations.

Fragment #	P-indexer	NoP Indexer Mix	Indexed Size	Gene Region	Gene Function
5040	GCGC	GCCN	416	*yehX*	putative ATP-binding component of a transport system
8349	GCGC	GCCN	526	*yhjX*	putative resistance protein
8511	GCGC	GCCN	726	*mtlD*	control region for mannitol-1-phosphate dehydrogenase
9984	GCGC	GCGN	884	*ytfM*	ORF, hypothetical protein
2465	GCGC	GCTN	479	*putP*	major sodium/proline symporter
9550	GCGC	GCTN	797	*yjbH*	ORF, hypothetical protein
9343	GCGC	GCTN	1064	*katG*	catalase; hydroperoxidase HPI(I)
399	GCGC	TAAN	322	*fhuA*	outer membrane protein receptor for ferrichrome and colicin M
9696	GCGC	TAAN	560	*rpiR*	transcriptional repressor of *rpiB* expression
9983	GCGC	TAAN	1263	*msrA*	control region for peptide methionine sulfoxide reductase
2002	GCGC	TACN	384	*mipB*	putative transaldolase
2337	GCGC	TACN	1832	*b0960*	ORF, hypothetical protein
9344	GCGC	TCTN	306	*yijE*	ORF, hypothetical protein
8187	GCGC	TGTN	615	*prlC*	oligopeptidase A
5869	GCGC	TTGN	466	*hyfR*	putative 2-component regulator, interaction with σ54
10402	GCGC	TTGN	592	*deoA*	thymidine phosphorylase

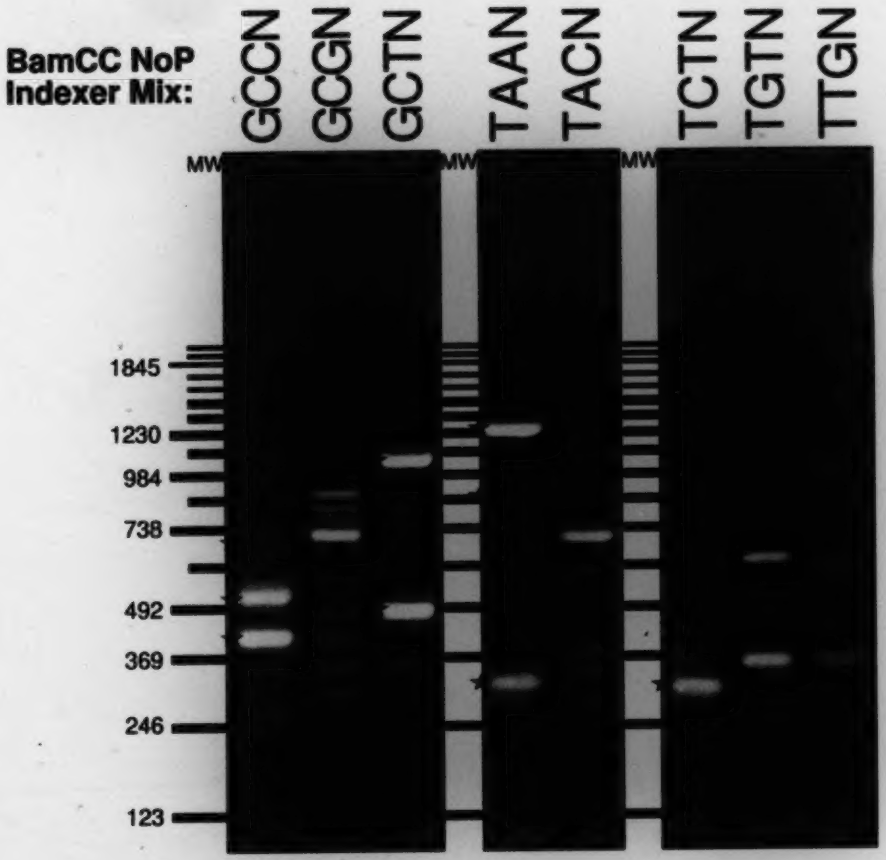


FIGURE 3.6: Correlation between predicted and detected indexing target fragments.
To confirm the ability of the EcoliDB database to identify *Fok*I restriction fragments which could be targeted by a particular indexer combination and amplified, the database was used to identify eight NoP indexer mixes were identified that were predicted to target specific *Fok*I fragments in ligations using P-GCGCxBam as the P-indexer. The number and indexed length of the fragments targeted by each NoP indexer mix were identified, enabling comparison of predicted and observed results.
Eight standard indexing ligations, each containing P-GCGCxBam and the appropriate NoP indexer mix, were assembled. Ligation was performed using the standard IGP benchmark ligation conditions, followed by PCR and amplified product analysis by agarose gel electrophoresis. Correctly indexed and amplified target fragments predicted by the EcoliDB database are denoted with a red star. Columns of black bars marked "MW" represent 123-bp ladder molecular weight markers.

number of other products of sizes not predicted by the database. Repetition of this experiment with ligations containing a second *Fok*I digest of independently-purified *E. coli* chromosomal DNA demonstrated the pattern of amplification product generation to be highly reproducible both for predicted and "unanticipated" indexed amplicons (data not shown). The reproducibility of the amplified product patterns suggests that the majority of the unanticipated amplicons are legitimate fragments generated by aberrant *Fok*I digestion patterns as described above, and that any misligation taking place is highly specific for particular mismatched cohesive end sequences. In some instances these "unanticipated" amplification products were more efficiently amplified from the complex IGP ligation mixtures than were the predicted target fragments. Given the complexity of the targeted DNA digest, and of the combination of indexers, amplification of products other than those specifically predicted by the database was not unexpected, and the reproducibility of the pattern was amenable to effective exploitation for indexed genomic profiling studies. The successful amplification of each target fragment predicted in the database demonstrated the utility of EcoliDB in the development of indexing-based molecular subtyping approaches using *E. coli* as the testbed for evaluation of IGP protocols.

3.3.7 Indexed genomic profiling of *E. coli* K12 strain MG1655

A genomic profile of *E. coli* K12 strain MG1655 was generated by indexing a *Fok*I digest of *E. coli* MG1655 chromosomal DNA with the P-indexer P-GCGCxBam and each of the 64 NoP indexer mixes (FIGURE 3.7). The survey of 1/256th of the indexing information present in the *E. coli* genome was performed using standard IGP protocols for ligation, PCR and agarose gel electrophoresis. The large number of amplified products presented an appropriate information density across the 64 indexing ligation and amplification reactions, providing a high number of discrete informative reference points contributing to the discriminatory power of the IGP approach. The pattern of amplified indexed products was reproducible for separate *E. coli* MG1655 genomic DNA preparations and *Fok*I digests (data not shown). Of the 105 indexing targets greater than 100 bp in length predicted from the EcoliDB database, 78 were

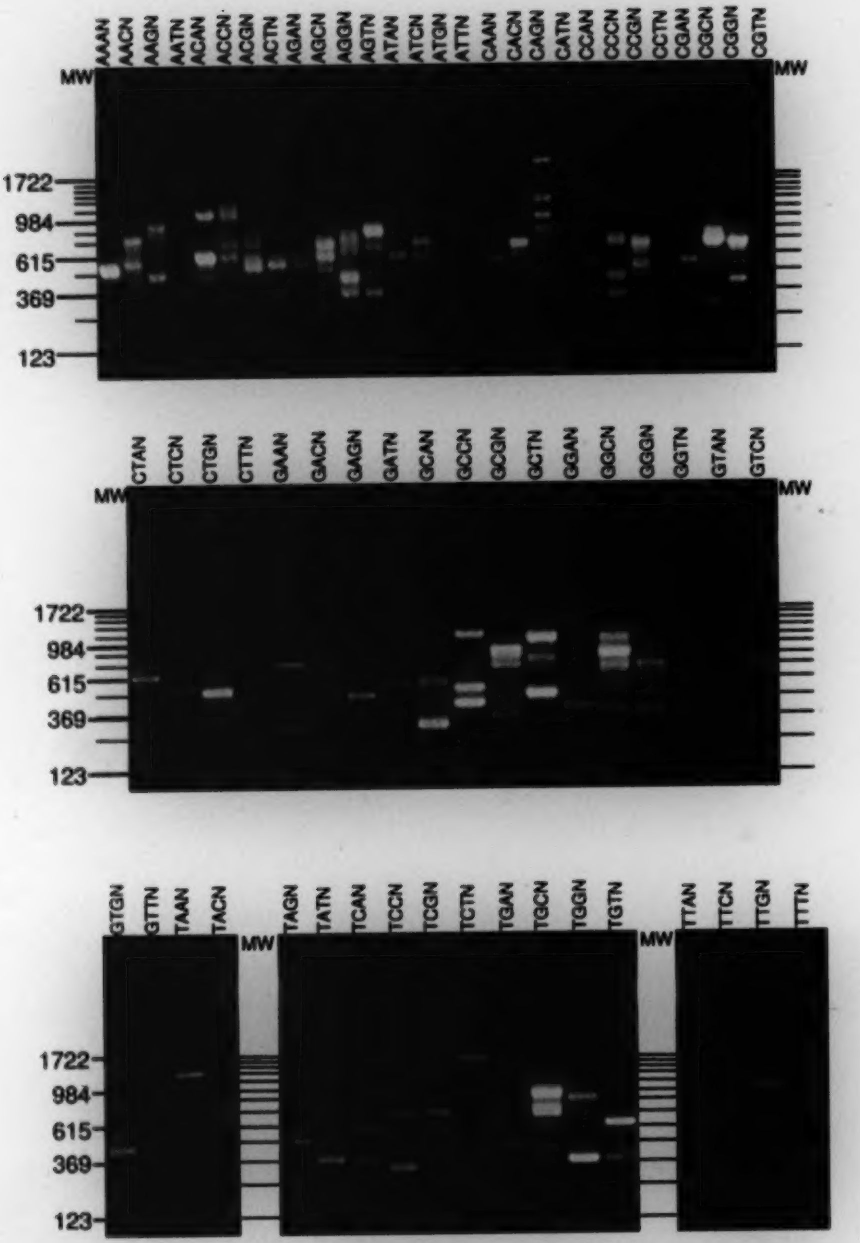

AAAN
AACN
AAGN
AATN
ACAN
ACCN
ACGN
ACTN
AGAN
AGCN
AGGN
AGTN
ATAN
ATCN
ATGN
ATTN
CAAN
CACN
CAGN
CATN
CCAN
CCCN
CCGN
CCTN
CGAN
CGCN
CGGN
CGTN
MW
MW
1722
984
615
369
123
CTAN
CTCN
CTGN
CTTN
GAAN
GACN
GAGN
GATN
GCAN
GCCN
GCGN
GCTN
GGAN
GGCN
GGGN
GGTN
GTAN
GTCN
MW
MW
1722
984
615
369
123
GTGN
GTTN
TAAN
TACN
MW
TAGN
TATN
TCAN
TCCN
TCGN
TCTN
TGAN
TGCN
TGGN
TGTN
MW
TTAN
TTCN
TTGN
TTTN
1722
984
615
369
123

FIGURE 3.7: Indexed genomic profiling of *E. coli* K12 strain MG1655.
A genomic profile of *E. coli* K12 strain MG1655 was generated by indexing a *Fok*I digest of *E. coli* MG1655 chromosomal DNA with the P-indexer P-GCGCxBam and each of the 64 NoP indexer mixes. Standard IGP protocols for ligation, PCR and agarose gel electrophoresis were employed. Columns of black bars marked "MW" represent 123-bp ladder molecular weight markers.

indexed and amplified. Of the 91 fragments greater than 200 bp in length predicted in the database, 77 fragments were indexed and amplified, or 85% of the targeted fragments predicted by the database for this series of indexer combinations and size range. The amplification by DNA indexing of a targeted restriction fragment subpopulation, presenting an objective sample of the indexing information content present in the genome, provided an information density and level of discriminatory power anticipated to be appropriate for high-resolution comparison with closely-related bacterial strains by indexed genomic profiling.

3.3.8 Differentiation of *E. coli* strains by IGP

3.3.8.1 Comparison of four E. coli *strains using a small subset of IGP reactions*

A simple experiment demonstrating the ability of IGP studies to differentiate between related bacterial strains using small numbers of indexer combinations was performed. In this experiment a small fraction (1/2056) of the total number of indexing fragment classes were targeted from *Fok*I digests of the genomic DNA of four *E. coli* laboratory strains: MG1655, JM109, JM110 and W3110. For each strain, four indexing ligations were assembled using the P-indexer P-CGCGxBam and the NoP indexer mixes OH-GCANxBamCC, OH-GCCNxBamCC, OH-TAGNxBamCC, and OH-TGGNxBamCC. From the EcoliDB indexing database for the *E. coli* MG1655 genomic DNA sequence, it was predicted that ligations containing P-CGCGxBam and OH-GCANxBamCC would generate an 808-bp amplicon and a 430-bp amplicon from strain MG1655 DNA. Similarly, ligations containing P-indexer and OH-GCCNxBamCC were anticipated to target a 1232-bp indexed fragment and a 389-bp fragment. Reactions employing P-indexer and OH-TGGNxBamCC were predicted to generate two amplicons of 886 bp and 349 bp, respectively. No targets for P-indexer and OH-TAGNxBamCC could be predicted from the published genomic DNA sequence for *E. coli* MG1655. Standard IGP protocols for ligation, PCR and agarose gel electrophoresis were followed.

Even such a small subset of indexer combinations as those used in this IGP "snapshot" of several laboratory strains of *E. coli* presented sufficient discriminatory power for strain differentiation while demonstrating the significant level of relatedness between the genomic restriction fragment populations (FIGURE 3.8). The 430-bp amplicon targeted by the cohesive end sequence mixture GCAN was amplified from strains MG1655 and W3110, but was not present in strains JM109 and JM110, while the 808-bp amplicon was present in all four strains. Both the 1232-bp and 389-bp amplicons predicted for the GCCN NoP indexer mix were present in all four strains. For the TGGN NoP indexer mix, the predicted 886-bp and 349-bp amplicons were targeted in three strains but not in strain JM110. Also, several unanticipated amplicons common to DNA digests derived from strains MG1655 and W3110 were not identified in digests of JM109 or JM110. Together with the evidence provided by the indexing in all tested strains of a 1.2 kb amplicon for the cohesive end sequence mixture TAGN, for which no predicted indexable fragment was identified from the strain MG1655 sequence, these data establish that even unanticipated indexed fragments may be reproducibly amplified and effectively utilized to further enhance the discriminatory power of indexed genomic profiling. Small subsets of indexing fragment classes are sufficient to discriminate among and demonstrate relatedness between similar bacterial strains.

3.3.8.2 Indexed genomic profiling of three E. coli *laboratory strains*

Molecular subtyping of three common laboratory strains of *E. coli* was performed by indexed genomic profiling. For each strain, a set of 28 indexing reactions employing a single P-indexer (P-AATGxBam) and one NoP indexer mix per reaction was assembled. Standard protocols for IGP ligation, PCR and agarose gel electrophoresis were followed for this experiment. A characteristic pattern of amplification products was observed across the 28 IGF reactions for each strain (FIGURE 3.9). Over 160 discrete profile datapoints, each presenting specific fragment size and cohesive end sequence information, were useful for discriminating strain identity. The ability of indexed genomic profiling to differentiate between closely

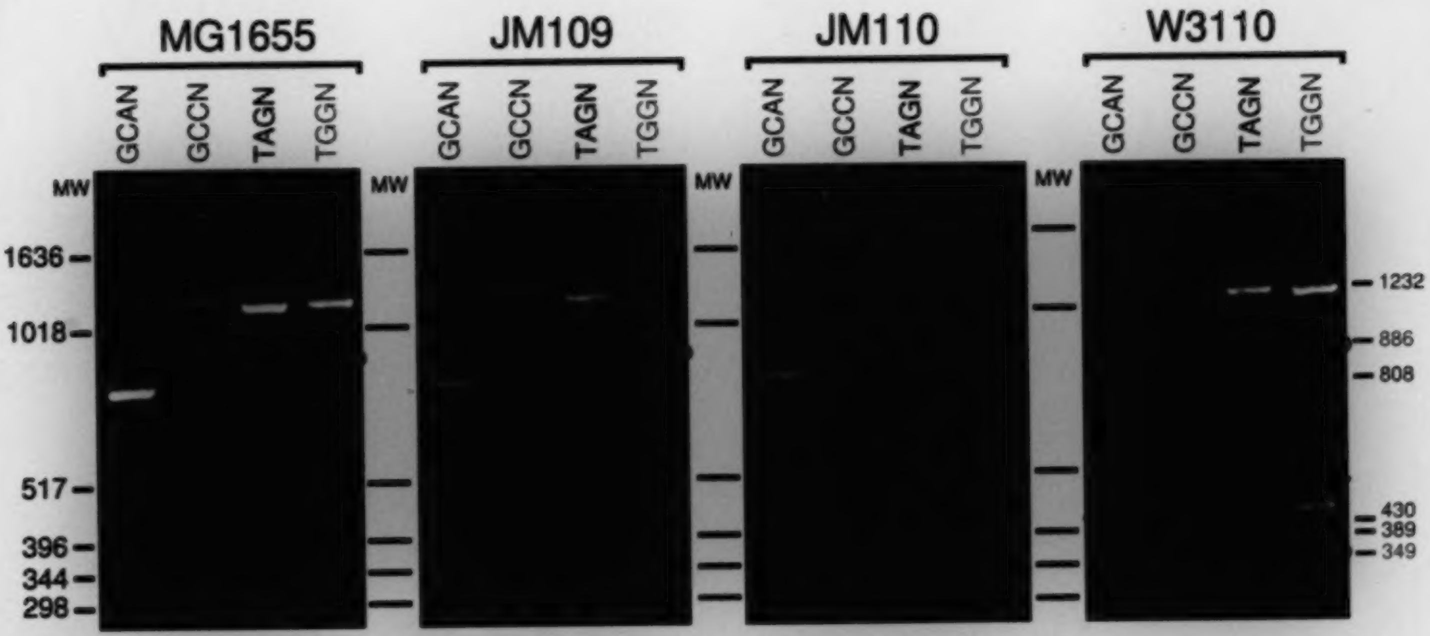


FIGURE 3.8: Comparison of four *E. coli* strains using a small subset of IGP reactions.
A small fraction (1/2056) of the total number of indexing fragment classes were targeted from *FokI* digests of the genomic DNA of *E. coli* laboratory strains MG1655, JM109, JM110 and W3110, in order to demonstrate the ability of IGP studies to differentiate between related bacterial strains using small numbers of indexer combinations. For each strain, four standard IGP indexing ligations were assembled using the P-indexer P-CGCGxBam and the NoP indexer mixes OH-GCANxBamCC, OH-GCCNxBamCC, OH-TAGNxBamCC, and OH-TGGNxBamCC. From the EcoliDB indexing database for the *E. coli* MG1655 genomic DNA sequence, it was predicted that ligations containing P-CGCGxBam and OH-GCANxBamCC would generate an 808-bp amplicon and a 430-bp amplicon from strain MG1655 DNA (noted in red). Ligations containing P-indexer and OH-GCCNxBamCC were anticipated to target a 1232-bp indexed fragment and a 389-bp fragment (noted in blue). Reactions employing P-indexer and OH-TGGNxBamCC were predicted to generate two amplicons of 886 bp and 349 bp (noted in green). No targets for P-indexer and OH-TAGNxBamCC could be predicted from the published genomic DNA sequence for *E. coli* MG1655. Standard IGP protocols for ligation, PCR and agarose gel electrophoresis were used. Columns of black bars marked "MW" represent Gibco-BRL's 1-kb ladder molecular weight markers.

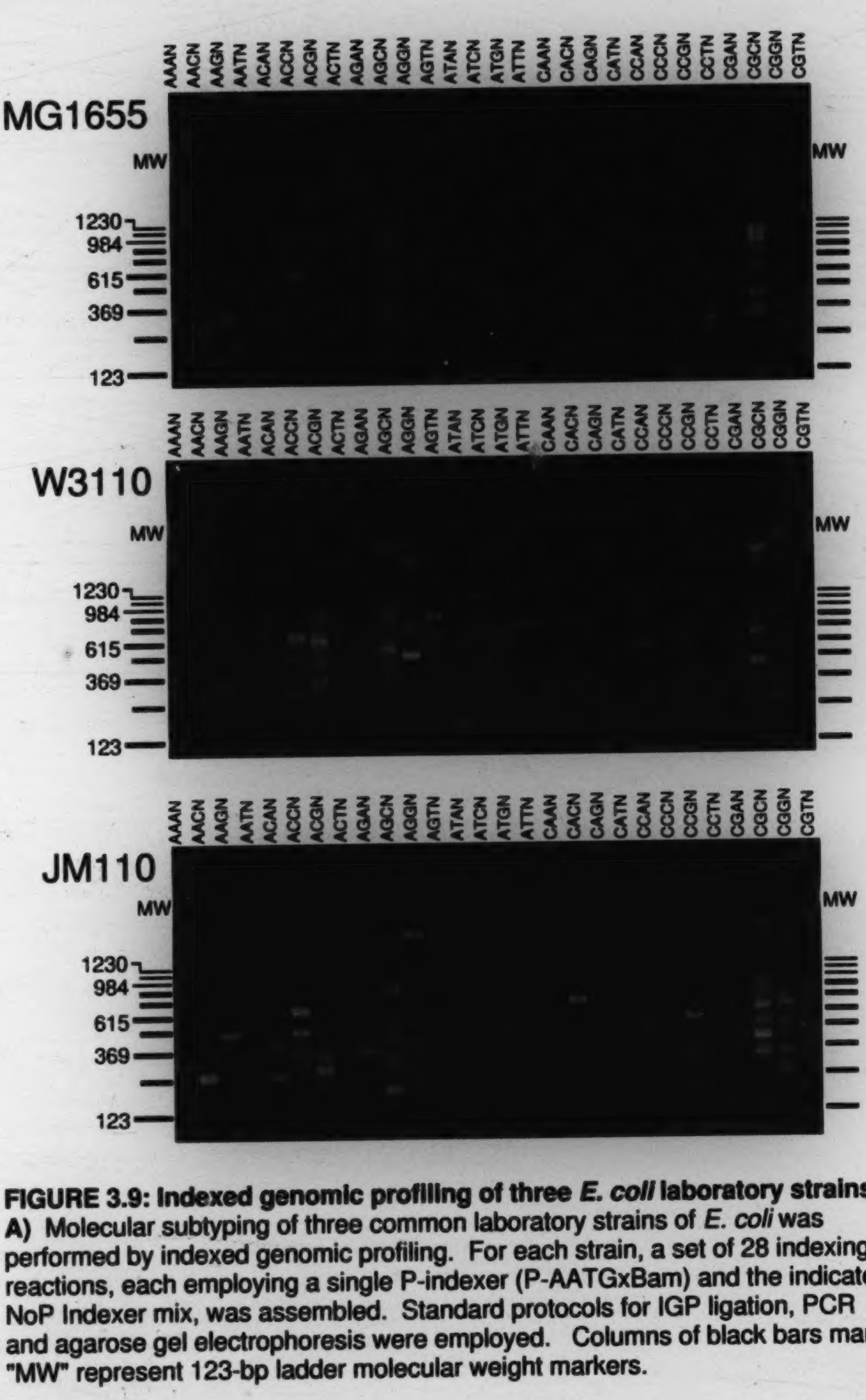

MG1655
W3110
JM110
MW
MW
1230
984
615
369
123
AAAN
AACN
AAGN
AATN
ACAN
ACCN
ACGN
ACTN
AGAN
AGCN
AGGN
AGTN
ATAN
ATCN
ATGN
ATTN
CAAN
CACN
CAGN
CATN
CCAN
CCCN
CCGN
CCTN
CGAN
CGCN
CGGN
CGTN

FIGURE 3.9: Indexed genomic profiling of three *E. coli* laboratory strains. **A)** Molecular subtyping of three common laboratory strains of *E. coli* was performed by indexed genomic profiling. For each strain, a set of 28 indexing reactions, each employing a single P-indexer (P-AATGxBam) and the indicated NoP Indexer mix, was assembled. Standard protocols for IGP ligation, PCR and agarose gel electrophoresis were employed. Columns of black bars marked "MW" represent 123-bp ladder molecular weight markers.

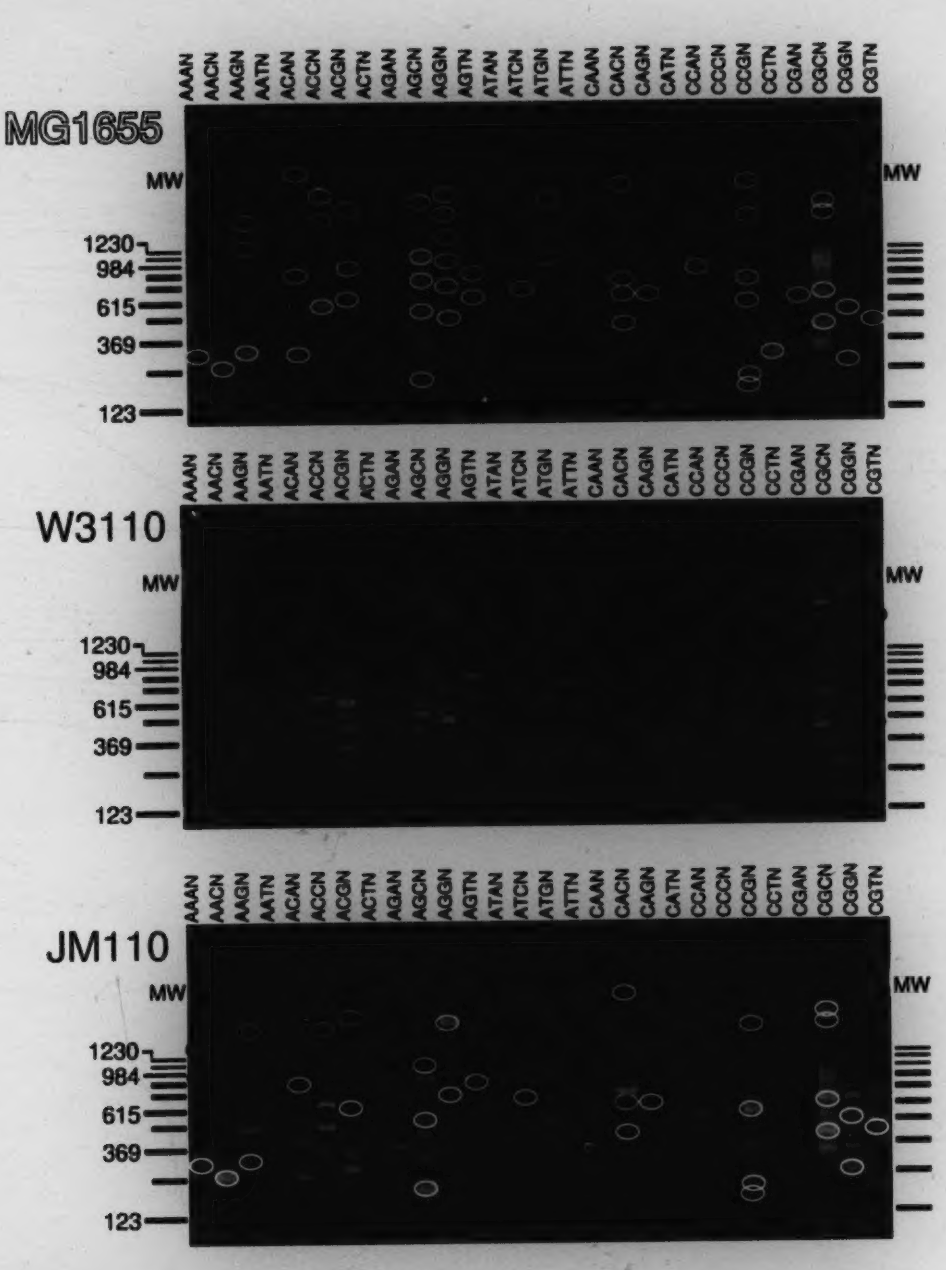

MG1655
W3110
JM110
AAAN
AACN
AAGN
AATN
ACAN
ACCN
ACGN
ACTN
AGAN
AGCN
AGGN
AGTN
ATAN
ATCN
ATGN
ATTN
CAAN
CACN
CAGN
CATN
CCAN
CCCN
CCGN
CCTN
CGAN
CGCN
CGGN
CGTN
MW
1230
984
615
369
123

FIGURE 3.9: Indexed genomic profiling of three *E. coli* laboratory strains. B) Annotated IGP profile of three *E. coli* strains highlighting strain profile similarities and differences. Selected strain MG1655 profile bands are circled in yellow, strain W3110 profile bands are circled in blue, and strain JM110 profile bands are circled in red. Bands common to MG1655 and W3110 are noted in green; bands common to MG1655 and JM110 are noted in orange; bands common to W3110 and JM110 are noted in purple; and bands common to all three *E. coli* strains are noted in white.

related strains of the same bacterial species was effectively demonstrated.

3.3.8.3 Indexed genomic profiling of Staphylococcus *reference species and clinical isolates*

Indexed profiles of clinical isolates of *S. aureus* and *S. epidermidis* and a reference strain of *S. lugdunensis* were generated employing the protocols for ligation and PCR developed for IGP applications using *Fok*I digests of *E. coli* genomic DNA. For each *Staphylococcus* species, a set of indexing ligations was assembled containing the same combinations of P-indexer and NoP indexer mixes employed for the generation of profiles for *E. coli* strains. Each species presented a profile that was highly specific (FIGURE 3.10). This experiment demonstrated that IGP provides a simple method of discriminating between related bacterial species present in clinical isolates or reference cultures.

Prior determination of genomic sequence data was not necessary for the generation of specific *Staphylococcus* species profiles, presenting information regarding fragment size and cohesive end sequence for hundreds of profile datapoints across the species profiled. If desired, any informative datapoint could be used to provide sequence data regarding the species from which it was generated. Isolation of particular amplified products from agarose gel, followed by direct cycle sequencing of the indexed fragment using the BamCC directional sequencing primer, would provide underlying sequence data for the chromosomal region from which the targeted restriction fragment was derived.

3.4 DISCUSSION

3.4.1 Summary of indexed genomic profiling studies

An indexing-based approach to microbial molecular subtyping was developed and demonstrated by adapting existing indexing protocols for the complexity of microbial genome analysis. The use of pools of NoP indexer sequences in ligations of bacterial genomic digests provided an efficient method of indexing for profiling applications. Ligation conditions for bacterial profiling were established through the

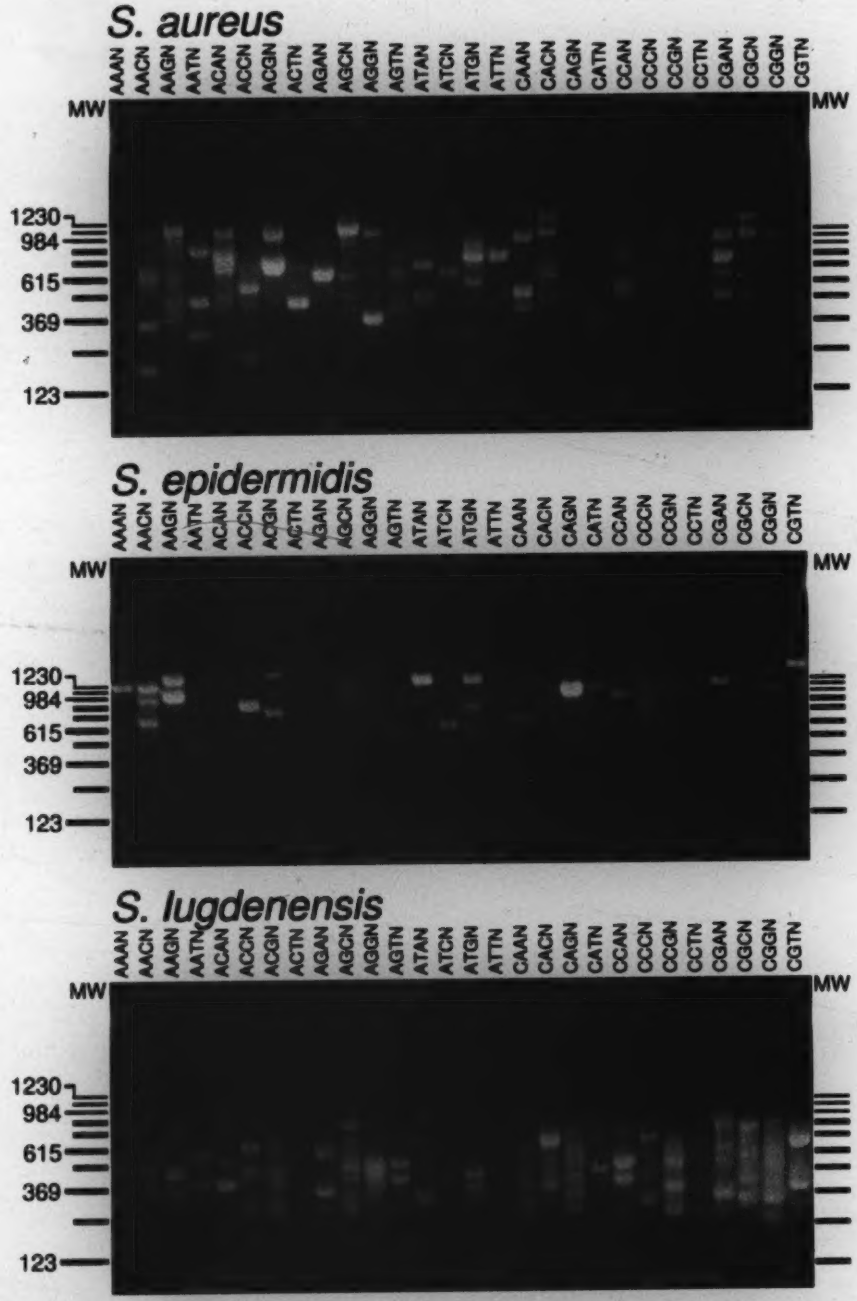

S. aureus
AAAN AACN AAGN AATN ACAN ACCN ACGN ACTN AGAN AGCN AGGN AGTN ATAN ATCN ATGN ATTN CAAN CACN CAGN CATN CCAN CCCN CCGN CCTN CGAN CGCN CGGN CGTN
MW
MW
1230
984
615
369
123
S. epidermidis
AAAN AACN AAGN AATN ACAN ACCN ACGN ACTN AGAN AGCN AGGN AGTN ATAN ATCN ATGN ATTN CAAN CACN CAGN CATN CCAN CCCN CCGN CCTN CGAN CGCN CGGN CGTN
MW
MW
1230
984
615
369
123
S. lugdenensis
AAAN AACN AAGN AATN ACAN ACCN ACGN ACTN AGAN AGCN AGGN AGTN ATAN ATCN ATGN ATTN CAAN CACN CAGN CATN CCAN CCCN CCGN CCTN CGAN CGCN CGGN CGTN
MW
MW
1230
984
615
369
123

FIGURE 3.10: Indexed genomic profiling of *Staphylococcus* species.
Indexed profiles of *S. aureus*, *S. epidermidis* and *S. lugdunensis* were generated employing the protocols for ligation and PCR developed for IGP applications using *Fok*I digests of *E. coli* genomic DNA. For each *Staphylococcus* species, a set of 28 indexing ligations, each employing a single P-indexer (P-AATGxBam) and the indicated NoP indexer mix, was assembled. Standard protocols for IGP ligation, PCR and agarose gel electrophoresis were employed. Columns of black bars marked "MW" represent 123-bp ladder molecular weight markers.

evaluation of indexer concentration, DNA concentration, and ligase concentration. *Taq* DNA ligase and *E. coli* DNA ligase were evaluated as IGP indexing ligases against T4 DNA ligase. Further development of reaction conditions for IGP ligations employing T4 DNA ligase established reaction parameters including incubation temperature, incubation time and ligase concentration. Software was developed to facilitate manipulation of bacterial genome sequence data for DNA indexing analysis. Initial application of the modified protocols to the molecular fingerprinting and differentiation of several *E. coli* strains was accompanied by predictive modelling based on the published genomic DNA sequence of *E. coli* strain MG1655. The ability of IGP studies to differentiate between related bacterial strains using small numbers of indexer combinations was demonstrated. Ligations generating legitimate indexed amplicons from predicted *Fok*I target fragments were amplified without substantial levels of background amplification products. Molecular subtyping of three common laboratory strains of *E. coli* was performed by indexed genomic profiling using complete sets of NoP indexer mixes. Indexed genomic profiles were generated from clinical isolates and reference strains of several *Staphylococcus* species. Prior determination of genomic sequence data was not necessary for the generation of specific *Staphylococcus* species profiles, presenting information regarding fragment size and cohesive end sequence for hundreds of profile datapoints across the species profiled. Indexed genomic profiling provides excellent discriminatory power in the form of an information-dense molecular fingerprint derived by objective sampling of microbial genetic structure.

3.4.2 Indexed genomic profiling as a potentially definitive method for bacterial genotyping

Indexed genomic profiling is a powerful molecular genotyping technique that permits microbial species and strain differentiation through the comparison of reproducible discriminatory amplification product patterns generated from indexed genomic restriction fragments. The profiles obtained by this approach are the result of simultaneous targeting of multiple loci by specific indexer combinations and represent an objective sample of the entire genome, rather than a specific set of genes or operons as for techniques like ribotyping or RFLP. Relatively small amounts of Type IIS endonuclease-digested genomic DNA are required to perform this technique. Application of IGP to clinical isolates of uncharacterized strains is not dependent on prior knowledge of the underlying DNA sequence. In instances in which the genomic sequence, or some interesting portion of it, has been determined, predictive evaluation of anticipated IGP results is possible and may provide the option of a more targeted experimental design. Its exquisite discriminatory power, widespread applicability, excellent reproducibility and information density make IGP a potentially definitive genotyping method for bacterial species and strain differentiation.

The discriminatory power of IGP, more so than any other molecular typing technique currently available, approaches the ideal of the "gold standard" for bacterial species and strain differentiation. Claims for another technique based on the amplification of restriction fragment subpopulations from genomic digests, AFLP, suggest that the method, based on classic Type II endonuclease cleavage and cohesive end sequences, offers a 10- to 50-fold increase in useful data points over RFLP analysis [150]. Due to the exploitation of informative cohesive end sequences by IGP, the information-rich data of an indexed genomic profile exceeds the information density of AFLP fingerprints by at least two orders of magnitude. IGP analysis provides comparable reproducibility with AFLP approaches, while permitting access to a wider range of chromosomal locations (all of them, technically) for evidence of genetic variability. As a result of the potential for complete access to and description of genetic structure, IGP studies may be tailor-made to include the simultaneous targeting of specific gene regions of interest within the generalized profile (such as the detection of the shiga toxin gene among *E. coli* strains within a molecular subtyping experiment, for example.) AFLP is unable to provide this level of flexibility. Taken together, these factors provide IGP approaches with unprecedented discriminatory power unmatched even by AFLP methods. The potential for sequence determination of strain-specific amplification products using the BamCC directional primer following gel isolation of the fragment of interest adds a further level of discriminatory power to IGP studies of uncharacterized clinical isolates. IGP analysis is amenable to automation, actualizing the potential for high throughput using this technique. Minor disadvantages in the cost and technical skill required to utilize IGP effectively in the context of a clinical laboratory may be partially ameliorated by protocol modification, and are outweighed by the enhanced discriminatory power and reproducibility provided by the IGP technique.

3.4.3 Potential applications of indexed genomic profiling

3.4.3.1 Clinical and epidemiological studies of pathogenic Staphylococcus *species and strains*

Widespread in nature, the *Staphylococci* are gram-positive nonmotile cocci that may behave as commensals or as pathogens in humans. The opportunistic pathogenicity of *S. aureus* as a source of acute pyogenic infection in humans has been well documented. The rise of hospital-acquired ("nosocomial") infections of methicillin-resistant *S. aureus* (MRSA) is of world-wide clinical and epidemiological importance [161, 162]. *S. epidermidis* has emerged as a major pathogen associated with infections in patients with implanted foreign devices [163]. *S. lugdunensis* is also a common cause of nosocomial infection, often colonizing catheters or prosthetic devices [164]. The development of methods capable of discriminating between staphylococcal species, and between pathogenic strains and those capable of asymptomatic colonization, is vital for analysis of pathogenesis, virulence gene distribution and the epidemiology of antibiotic resistance.

The ability of molecular typing systems to distinguish among epidemiologically unrelated isolates is related to the genetic variation seen in the chromosomal DNA of a bacterial species and to the level of sensitivity and discrimination of that variation by the typing method employed [165]. Recent comparative genomic studies have revealed that extensive variation in gene content is exhibited among strains of many important pathogenic bacterial species. Among *E. coli* strains, as many as 25% of genes have been reported to be strain-specific [166], while 22% of *H. pylori* genes were demonstrated to be nonessential to basic cellular processes [167]. These data suggest that a surprisingly large fraction of the genomes of these human pathogens are devoted to contingency functions, suggesting a mechanism for generation and maintenance of interstrain genetic variation [168]. The clinical isolates most commonly identified with infections thus represent only a small fraction of the many strains that constitute a species [169]. Further, this subset may exhibit relatively little internal genetic diversity, with the result that discriminating strain specificity among clinical isolates can be difficult with current molecular techniques [137].

Epidemiological and clinical studies of methicillin-resistant *S. aureus* illustrate the implications of significant genetic variation within strains of pathogenic bacteria. The genetic variation exhibited in staphylococcal species is highly significant, with ~22% of the *S. aureus* genome comprised of dispensable genetic material. 18 large regions of difference were identified,10 of which contained genes that encode putative virulence factors or proteins mediating antibiotic resistance [169]. Investigations of molecular population genetics have shown that relatively few MRSA clones are responsible for the majority of infections [169, 170], and that only a limited number of strain types can be discriminated by most typing approaches. While phenotypic methods are generally incapable of resolving MRSA isolates except into a few broad groups, methods that are sensitive to variation in genome content are capable of differentiating endemic from epidemic strains [132].

Indexed genomic profiling is ideally suited for bacterial typing applications of this nature. The capacity for objective high-resolution sampling of a large number of chromosomal regions simultaneously, coupled with the ability to specifically target putative virulence or antibiotic-resistance gene regions, provides the clinical microbiologist with a molecular fingerprint with specific information regarding strain pathogenicity. Closely-related clones which present different clinical characteristics may be resolved by increasing the number of indexing fragment classes surveyed, and thus increase the discriminatory power of the specific investigation to the required level. Finally, the reproducibility and potential for high-throughput application of IGP facilitates interlaboratory comparison and library cataloging of strain-specific IGP characteristics.

3.4.3.2 *Studies of microbial community diversity*

The number and diversity of microbial species colonizing a particular environment is typically enormous. The complex interactions between the species present in a microbial community are of great scientific interest and, in many cases, economic or environmental importance. Microbial communities in soil samples, marine sediments, biofilms, and even in dental plaque have been investigated using a wide range of culture-based and culture-independent techniques. However, currently available analytical techniques are incapable of providing complete information regarding the identities of constituent species in microbial communities [171]. Ideal methods for microbial community analysis would provide reliable, objective species identification permitting high resolution monitoring of community diversity, dynamics and stability [172].

Microbial ecologists have traditionally depended on culture-based methods for analysis of species diversity and community dynamics. However, the fraction of species cultured from microbial communities has been demonstrated to be typically less than 1% of the total [173, 174]. Although DNA-based analytical techniques avoid the bias associated with culturing of bacterial species from complex communities, the number and scope of culture-independent methods effective for community analysis is still quite limited [175]. The three methods currently in vogue are ribotyping, terminal restriction fragment length polymorphism (T-RFLP), and RAPD analysis [176]. Inadequacies with each of the approaches continue to hamper research efforts to describe microbial community structure in a taxonomically precise manner.

The most common culture-independent method currently used to analyze the species composition of microbial communities is ribotyping. Two versions of this technique have been employed: the use of rRNA-targeted oligonucleotide probes [177] and amplified ribosomal DNA restriction analysis (ARDRA) [178]. The 16S rRNA targeted by either method may be too highly conserved to discriminate between different species with almost identical sequences. The lack of a complete *a priori* description of rRNA diversity may allow probes designed to be specific for particular species to hybridize to the rRNA sequences of as yet uncharacterized species, with the consequence that those species are misidentified [172]. Probing approaches are not easily automated, constraining the number of samples that can be processed with a limited set of probes [172]. As ARDRA involves the restriction endonucleolytic digestion of PCR-amplified 16S rDNA sequences followed by resolution of restriction fragments by agarose or polyacrylamide gel electrophoresis, this method's main limitation lies in the choice of restriction enzymes, critical for optimum product resolution [176]. For each microbial community studied, preliminary tests for appropriate enzyme choice must be performed, reducing the comparability of results between samples [175]. Little information regarding specific ribosomal DNA restriction patterns for particular species can be generated by this method [176].

T-RFLP analysis is considered the most powerful technique currently available for comparisons of species diversity in environmental samples [179]. T-RFLP is an alternative approach to ARDRA that employs fluorescently-labeled primers for detection of terminal fragments of a digested 16S rDNA amplicons. The variation in restriction site location along these amplicons provides a reproducible high-resolution fingerprint of community diversity which may compared to rRNA databases for comparative sequence analysis [180, 181]. The T-RFLP method has been applied to the differentiation of microbial communities and the identification of specific organisms within communities [182, 183], and its usage among microbial ecologists is increasing. However, recent studies have demonstrated T-RFLP to be ineffective in assessing relative phylotype richness and structure in highly complex soil communities [182]. Additionally, substantial levels of variability were observed among replicate profiles, raising serious questions regarding the validity of comparisons of microbial communities in different environmental samples using this technique [182].

In contrast to typing methods based on ribosomal nucleic acid sequence differences between species, fingerprinting of complete microbial communities by RAPD analysis compares the overall similarity of community composition [171]. For this whole-community approach, which avoids the requirement for specifically-designed PCR primers, the overview of community dynamics is disconnected from the ability to assay community diversity or to identify the constituent bacterial species. However, large numbers of degenerate or random primers are required to obtain enough data points for statistical comparison of community profiles. While other typing methods can provide information regarding the presence or absence of particular strains, no technique currently exists to completely characterize the structure of bacterial communities [171].

The potential application of indexed genomic profiling to studies of microbial community complexity is interesting to consider. IGP profiles would provide an objective sample of all fragments in a particular set of fragment classes from *Fok*I-digested DNA preparations representing all organisms in a microbial community, without relying on the genetic variability at a particular set of gene loci for species discrimination. The level to which the entire DNA population of the community is described may be regulated by the number of indexer combinations used to generate the profile. The reproducibility of IGP, the information density of each data point, and the potential ease of sequencing isolated indexed fragments of interest suggest that indexed community profiling (ICP) databases could be constructed to permit comparative sequence analysis and species identification. Indexing-based approaches may permit simultaneous community profiling and constituent species identification. Modifications of IGP protocols to employ individual NoP indexers would reduce the complexity and increase the interpretability of ICP studies. Use of fluorescently-labeled primers would permit ICP analysis using automated DNA sequencers, increasing the reproducibility and throughput of the method. Given the sensitivity of the new generation of automated DNA sequencing instrumentation, it could be feasible to omit the amplification step of IGP, and to ligate sets of four fluorescently-labeled P-indexers (labeled with spectrally-resolvable dyes) directly to *Fok*I digests of entire community DNA preparations, in order to generate quantitative profiles of bacterial communities. The potential of indexing as a tool for characterization and monitoring of microbial community diversity has yet to be explored.

4 Chapter IV: Global Gene Expression Profiling of *Saccharomyces cerevisiae* by 3'-end cDNA Indexing

4.1 INTRODUCTION

A major focus of recent large-scale biological research efforts has been the decoding and analysis of all genetic information present in several eukaryotic and numerous prokaryotic genomes. To achieve this goal, a complete view of a genome's physical structure must be combined with a thorough understanding of the functional characteristics inherent to this structure. As the focus of these projects shifts from large-scale nucleotide sequencing of chromosomes to comprehensive analysis of the functional genomics of these organisms, new techniques for functional analysis have been developed. In particular, technologies enabling the study of global gene expression profiles have been demonstrated to be powerful tools in the quest to discover physiological meaning behind genomic sequence.

The transcriptome, the set of mRNAs expressed by an organism or cell type at a given time under given conditions, is a major determinant of biochemical function and phenotype. In contrast to the genome, the transcriptome is dynamic and responsive to the program of normal physiological events, to environmental stimuli, and to perturbations due to disease processes [73, 184, 185]. Generation and comparison of global transcription profiles provide insight into gene function by identifying when, where, to what level and in response to what stimuli each particular gene is expressed, and may ultimately permit the elucidation of the complete transcriptional regulatory circuitry of an organism [186-188]. Several general approaches have been developed for the description of complete gene expression profiles, capable of characterizing and tracking changes in the transcriptome of the cell type or organism of interest.

Differential display

Differential display [189] has been widely used to compare expression levels between cell types or between cells subjected to various environmental cues, for relatively small subpopulations of genes [21]. Low-stringency amplification of cDNA tags by arbitrary primers, or by an arbitrary primer and a 3'-anchored poly(T) primer, generates a fingerprint on polyacrylamide gels, each of which is composed of a subpopulation of cDNAs sampled from the original complex mRNA source in a primer-dependent manner [190]. Differential display-based methods have several technical limitations, including imperfect reproducibility, the inability to efficiently survey the entire transcriptome, and poor sensitivity to rare mRNA species [159, 191, 192].

Serial analysis of gene expression

Serial analysis of gene expression (SAGE) [193] is designed to provide quantitative gene expression data for entire transcriptome characterization by concatenation, cloning and automated sequencing of short cDNA tags. The power of the approach has been demonstrated through studies of a wide range of cell types under a variety of physiological and pathological conditions [185, 194, 195]. However, this technically complex technique has several intrinsic shortcomings, including the requirement of large quantities (1 to >5 μg) of mRNA for each experiment; a protocol necessitating extensive cloning and large numbers of sequencing runs; incomplete transcriptome coverage in a single iteration of the protocol; difficulty in analyzing transcripts derived from uncharacterized gene sequences; the generation of internally redundant and nonspecific transcript-tag sequences; and a significant level of transcriptome profile inaccuracy due to sequencing errors [196-198]. An excellent study by Stollberg *et al.* identified and investigated theoretical and practical sources of error in SAGE experiments, and concluded that this method for global gene expression analysis significantly under-estimates the number of active genes and the fraction of genes expressed at low copy numbers [199].

cDNA microarrays

One of the most technologically advanced and powerful tools for global studies of gene expression is the cDNA microarray [188, 200, 201]. This approach involves the synthesis or attachment of nucleic acid sequences on glass slides at very high densities, over which fluorescently-labeled RNA or cDNA populations are washed. Transcripts are hybridized to complementary sequences following a massively-parallel search for a binding partner, and the level (relative or absolute) of fluorescence at each location of the affinity matrix is detected [188]. The power of this approach has been clearly demonstrated [194, 202-212]. However, at its present stage of development, even this technology is not without flaws. A non-linear relationship between signal strength and the amount of probe present can result in low precision of mRNA concentration estimates [213, 214]. More mundane considerations also affect the accessibility of microarray technology. Assembly of the microarrays themselves, in addition to the significant level of optical signal processing and data analysis software required for meaningful interpretation of results, is beyond the technical ability of most researchers, resulting in a dependence on commercial providers. As a consequence, research using this technique is limited to the analysis of cell types studied by economically-significant numbers of laboratories, except in the case of research groups with sufficiently large budgets to contract for custom microarray synthesis [213].

3'-end cDNA indexing

3'-end cDNA indexing was the first technique capable of describing entire mRNA populations for global gene expression studies [159]. Innately less susceptible to bias towards high-copy-number mRNAs than differential display, 3'-end cDNA indexing is capable of reporting the expression state of mRNA species at very low copy number in an mRNA population, which includes the vast majority of mRNAs expressed [159, 191]. Technically simpler than SAGE, it can be adapted to facilitate either relative or quantitative comparisons of expression level for each transcript surveyed [215, 216]. It presents an attractive and economical alternative to expression analysis by oligonucleotide array, particularly for investigations of global gene expression changes in organisms with as-yet-uncharacterized genomes.

To date, most applications of 3'-end cDNA indexing have focused on the differential expression of genes between tissues and through development for the same tissue. These studies have included the profiling of gene expression in murine cerebellum during development [216-218], comparisons of genes expressed differentially in murine duodenum and ileum [219], and identification of genes important in X-linked α thalassemia syndrome and in Alzheimer's disease [220, 221]. Several variants on 3'-end cDNA indexing concepts have been developed [98, 215, 222, 223], and a thorough evaluation of the ligation fidelity of cDNA indexing methods has been reported [223].

In this investigation, protocols for a modified approach to 3'-end cDNA indexing were developed. As a demonstration of the utility of 3'-end cDNA indexing for studies of transcriptome dynamics in response to environmental stimuli, these modified protocols were applied to the study of global gene expression changes in *Saccharomyces cerevisiae*.

4.1.1 *Saccharomyces cerevisiae* as a model organism for transcriptomics

Of immense economic importance as common baker's or brewer's yeast, *S. cerevisiae* is also one of the most important model organisms for scientific investigation of fundamental aspects of eukaryotic biology including the cell cycle, genome structure and maintenance, and gene regulation and metabolic circuitry. Dissection of the biology of this important organism has been facilitated by increasingly sophisticated molecular genetic tools. Elucidation of the complete genomic sequence of *S. cerevisiae* [224] has revealed our lack of understanding of yeast biology as much as it has increased the breadth of our knowledge regarding this organism. Less than half of the 6000 genes identified in *S. cerevisiae* have been "functionally characterized", and the dynamics of global gene expression are only beginning to be defined [225-227]. Estimates of the complexity of yeast transcriptome repertoires derived by several different methods indicate that only 4000-4500 ORFs are expressed under any given set of environmental conditions, with 80% of these transcript species expressed at low levels (0.1 to 2 transcripts per cell) [195, 226, 228]. A major focus in the post-genomic era of yeast research is the characterization of the network architecture forming the foundation of the entire regulatory circuitry of eukaryotic cells [186, 227, 229]. Development and application of a range of genomic, transcriptomic and proteomic approaches are necessary to achieve this goal. cDNA indexing is a technology potentially applicable to studies of gene expression dynamics in *S. cerevisiae* in response to environmental stimuli.

4.1.2 Global gene expression profiling of *S. cerevisiae* by 3'-end cDNA indexing

The modified 3'-end cDNA indexing approach developed in this investigation of global gene expression profiles in yeast is outlined in FIGURE 4.1. A set of 64 biotinylated indexer mixes is constructed by annealing a biotinylated primer (e.g. biotinylated BamCC primer) to each of a set of 64 nonphosphorylated indexing oligonucleotides complementary to all possible 4-base 5'-cohesive ends generated by Type IIS restriction enzymes. A complete poly(A)-tailed mRNA population is isolated from total cellular RNA obtained from each of several *S. cerevisiae* cultures grown under differing environmental conditions. The mRNA is reverse-transcribed into single-stranded cDNA using anchored poly(T) oligonucleotides to prime cDNA strand elongation, and second-strand cDNA is synthesized via RNA replacement (using RNAse H, *E. coli* DNA polymerase I and T4 DNA ligase). The ds-cDNA population, representative of the relative abundances of transcript species present in the initial mRNA population, is digested by a Type IIS restriction enzyme (e.g. *Fok*I). *Fok*I-digested cDNA is aliquoted into 64 ligation reactions, each containing one of the 64 biotinylated indexer mixes and DNA ligase. cDNA fragments bearing the cohesive end sequence complementary to that of one of the four indexers represented in the indexer mix are ligated to the particular indexer bearing that sequence. Biotinylated indexed cDNA fragments are captured using streptavidin-coated paramagnetic beads, washed to remove nonspecifically-bound DNA fragments, and amplified by PCR.

The two primers used to amplify 3'-end cDNA fragments are a (typically fluorescently-labeled) indexing primer (e.g. FAM-BamCC primer) acting as the 5' amplimer, and a single-base anchored oligo-d(T) primer as the 3' amplimer. In this way, a specific set of cDNA restriction fragments will be amplified in each of the 64

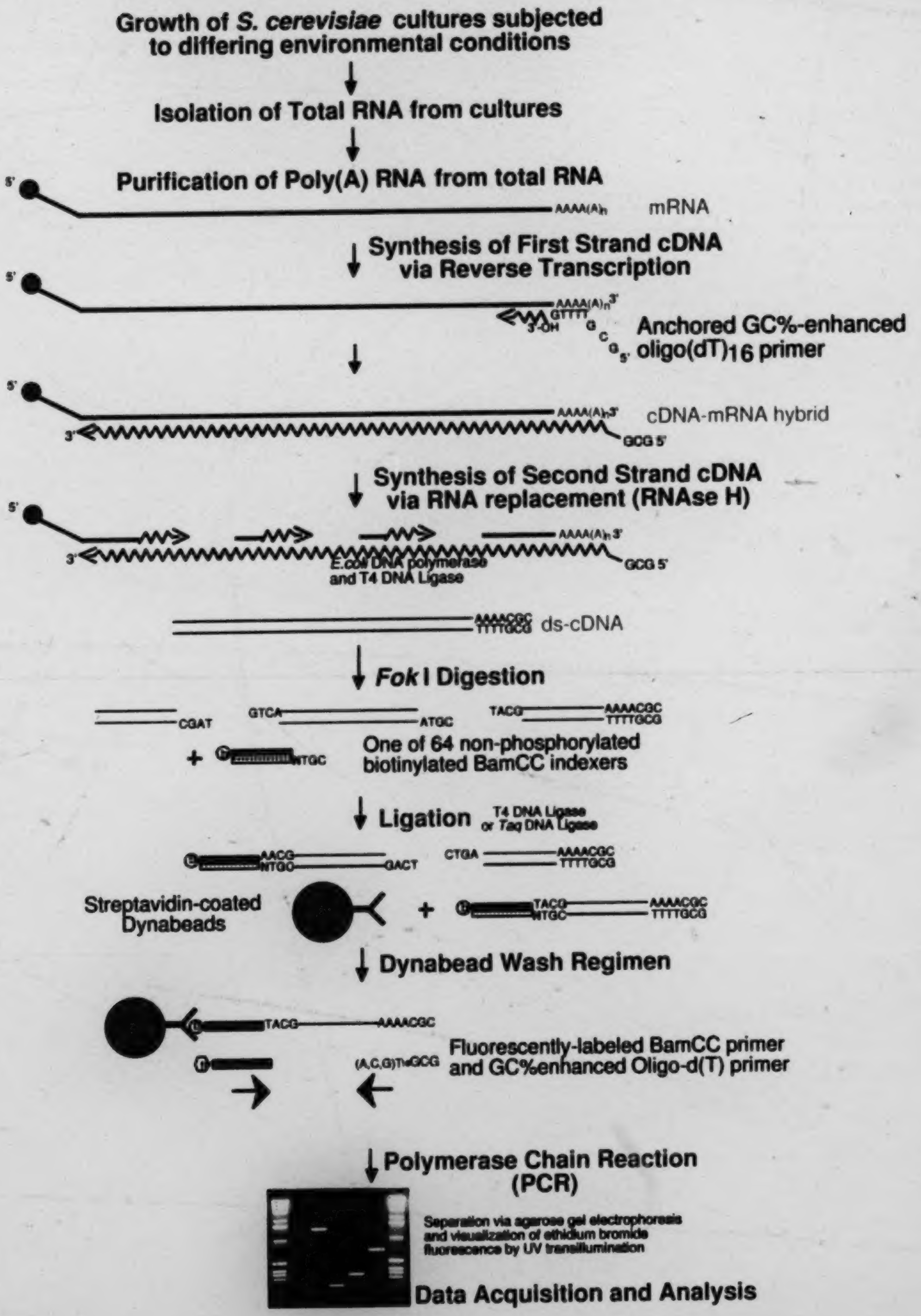

Growth of S. cerevisiae cultures subjected to differing environmental conditions
Isolation of Total RNA from cultures
Purification of Poly(A) RNA from total RNA
5'
AAAA(A)n
mRNA
Synthesis of First Strand cDNA via Reverse Transcription
5'
AAAA(A)n 3'
GTTTT
3'-OH
G C G 5'
Anchored GC%-enhanced oligo(dT)16 primer
5'
AAAA(A)n 3'
3'
GCG 5'
cDNA-mRNA hybrid
Synthesis of Second Strand cDNA via RNA replacement (RNAse H)
5'
AAAA(A)n 3'
3'
GCG 5'
E.coli DNA polymerase and T4 DNA Ligase
AAAACGC
TTTTGCG
ds-cDNA
Fok I Digestion
CGAT
GTCA
ATGC
TACG
AAAACGC
TTTTGCG
+
NTGC
One of 64 non-phosphorylated biotinylated BamCC indexers
Ligation
T4 DNA Ligase or Taq DNA Ligase
AACG
NTGC
GACT
CTGA
AAAACGC
TTTTGCG
Streptavidin-coated Dynabeads
+
TACG
NTGC
AAAACGC
TTTTGCG
Dynabead Wash Regimen
TACG
AAAACGC
(A,C,G)T16GCG
Fluorescently-labeled BamCC primer and GC%enhanced Oligo-d(T) primer
Polymerase Chain Reaction (PCR)
Separation via agarose gel electrophoresis and visualization of ethidium bromide fluorescence by UV transillumination
Data Acquisition and Analysis

FIGURE 4.1: Expression profiling of *S. cerevisiae* by 3'-end cDNA indexing.

reactions. Only those fragments bearing a poly(A) tail on one end and a *Fok*I cleavage site with the correct cohesive end on the other will be amplified in any particular reaction. (cDNA fragments with cohesive ends non-complementary to the particular indexer are not biotinylated. Fragments with the complementary cohesive end on both ends will not be amplified, due to dissociation of both BamCC primer-binding sites found on the non-phosphorylated indexer.)

This series of steps is repeated using other cDNA populations representing yeast grown under different environmental conditions or stimuli, using BamCC labeled by a different fluorescent dye for each experimental set. The result is several sets of 64 indexed and amplified 3'-end cDNA fragment subpopulations. The sample sets are pooled and analyzed by automated DNA sequencing instrumentation using fluorescence-based detection.

The genomic DNA sequence of *S. cerevisiae* encodes 6340 potential open reading frames (ORFs), 5885 of which are potential or identified protein-encoding genes [224]. Database queries identified 4559 ORF sequences giving rise to cDNA transcripts from which 3'-terminal *Fok*I restriction fragments exceeding 100 bp in length following indexer ligation and amplification may be generated. (This minimal length is easily sized by either agarose gel electrophoresis or by automated DNA sequencing instrumentation, and exceeds the size of most background amplification artifacts.) This indicates that analysis of gene expression in *S. cerevisiae* by 3'-end cDNA indexing using *Fok*I as the indexing restriction endonuclease would allow detection and analysis of 77.5% of protein-encoding genes. Similar database queries for 3'-terminal *Sfa*NI restriction fragments identified 4497 potential targets using the criteria described above, representing 76.5% coverage of *S. cerevisiae* ORFs. An indexing approach employing both *Fok*I and *Sfa*NI, and thus combining the ORF sets covered by each enzyme individually, would target 98.75% of the protein-encoding genes in fusion yeast. In other words, only 1.25% of yeast genes, or 74 protein-encoding ORFs, would not generate indexed 3'-terminal restriction fragments exceeding 100 bp in length following digestion of the corresponding ds-cDNA transcripts with either of these Type IIS endonucleases and attachment of indexers.

(For comparison, a survey of all published yeast microarray data to date has found that the transcripts of 144 yeast ORFs have never been detected under any experimental conditions using cDNA microarray analysis of gene expression [230]. Published gene expression studies in *S. cerevisiae* employing SAGE have only detected sequence tags corresponding to a total of 4665 yeast ORFs [195].) Each of the 74 genes refractive to 3'-end cDNA indexing using *Fok*I or *Sfa*NI are either effectively targeted using the same criteria by the Type IIS restriction enzyme *Bsm*AI, or are accessible by including indexed 3'-terminal fragments less than 100 bp in length in the analysis. The use of all three indexing endonucleases in this way provides the potential for complete coverage of the yeast transcriptome if required. cDNA indexing with any one of these enzymes alone is likely to provide an informative transcriptome profile for most experimental growth or stimulus conditions.

3'-end cDNA indexing using a single Type IIS endonuclease (for example, *Fok*I) divides the entire cDNA population into 64 non-overlapping subpopulations. If, following ligation of indexers, the cDNA in each of the 64 reactions is digested by two other Type IIS enzymes (in this case, *Sfa*NI and *Bsm*AI) prior to amplification, and if this process is repeated for each of the three indexing endonucleases, 192 non-overlapping subpopulations may be characterized. (This approach ensures that a cDNA fragment is only amplified if it carries a poly(A) tail at its 3'-end, and if it remains ligated to a complementary biotinylated indexer at its 3'-proximal restriction site for the case in which the enzyme producing the indexed cleavage site cuts closest, of the three enzymes employed in the experiment, to the poly(A) tail on that cDNA transcript.) If three discrete amplifications using the individual anchored poly(T) primers are used [i.e. poly$(T)_N$A, poly$(T)_N$C and poly$(T)_N$G] in place of a poly(T)-V primer mixture, 576 non-overlapping subpopulations are possible. This ability to subdivide mRNA populations markedly contrasts with differential display techniques in which subpopulations amplified by different sets of primers may theoretically overlap, and by which most genes would be displayed more than once when multiple sets of primers are used. It is this propensity for redundancy which limits conventional differential display techniques to the study of similar cell types - the fingerprinting pattern which would be generated by comparison of distantly related mRNA populations would be too complex for adequate analysis. The ability of indexing to subdivide mRNA populations also generates greater sensitivity for detection of rare mRNA species, even using conventional PCR.

4.1.3 ABSTRACT:

The purpose of this investigation was to develop modified 3'-end cDNA indexing protocols in order to facilitate global gene expression in *S. cerevisiae*. Double-stranded cDNA populations were synthesized from yeast total cellular mRNA in preparation for cDNA indexing analysis. The cDNA populations were demonstrated to be representative of the mRNA populations from which they were derived. *Fok*I digestion of cDNA populations generated indexable 3'-terminal cDNA fragments predicted from ORF sequence data of the *S. cerevisiae* genome. Indexers were ligated to the complementary cohesive end sequences of targeted *Fok*I-digested 3'-terminal cDNA fragments within a complex cDNA restriction digest. The indexed fragments were amplified using an indexing primer and transcript-specific primers. The selective capture of target fragments ligated to biotinylated indexers by streptavidin-coated paramagnetic beads was employed to reduce amplification reaction complexity, following the determination of stringent wash regimens eliminating nonspecific binding of nonbiotinylated cDNA. Using a series of artificial poly(A)-tailed indexable constructs, anchored GC-rich poly(T) primers were found to provide improved priming efficiency for cDNA population synthesis and for amplification relative to several other anchored poly(T) primer conformations. Artificial poly(A)-tailed indexable constructs were also used to evaluate the efficiency of 3'-terminal cDNA fragment amplification from templates bound to paramagnetic beads, to determine the amount of particular transcript species needed for target amplification following the ligation of NoP indexers, and to determine improved PCR cycling parameters. Ligation conditions providing high ligation fidelity with *Taq* DNA ligase were incorporated into the modified cDNA indexing protocol set.

Using the modified cDNA indexing protocols, differential gene expression profiles for yeast cultures exposed to various environmental stimuli were identified. Expression of the *GAL1* transcript was observed in yeast grown in galactose-containing medium, while no *GAL1* expression was detected in yeast grown in glucose-containing medium. Increased expression of the *BOP3* transcript was observed in cDNA populations derived from pheromone-treated yeast cultures relative to cDNA populations derived from untreated cultures. These findings were well-correlated with published data obtained by established methods of gene expression analysis. A limited survey of gene expression changes in yeast responding to saline shock performed using a small number of indexers generated results compatible with published data obtained in studies of yeast salt shock response using cDNA microarrays.

Analysis of 3'-end cDNA indexing data by automated fluorescence-based DNA sequencing instrumentation revealed the reproducibility of cDNA indexing profiles generated from independent parallel indexing ligations targeting individual cDNA populations and from distinct cDNA populations derived from parallel yeast cultures grown under identical conditions. Distinct indexed gene expression profiles were generated from cDNA populations derived from yeast cultures grown in the presence of differing environmental stimuli. Differences in the level of amplification of specific indexed 3'-terminal cDNA fragments were observed, indicating differences in the level of expression of specific mRNA transcripts between saline-treated and untreated yeast cultures. However, 3'-end cDNA indexing data sets were poorly correlated with data from published studies of saline shock response in *S. cerevisiae*. Unanticipated fragments were amplified, and certain anticipated indexed 3'-end cDNA fragments were not detected in 3'-end cDNA indexing data sets, indicating that refinement of the 3'-end cDNA indexing technique is necessary for effective application to global gene expression profiling in *S. cerevisiae*.

4.2 MATERIALS AND METHODS

4.2.1 *S. cerevisiae* strains and growth conditions

S. cerevisiae strain yAO6 is a STE12Δ deletion mutant of the common laboratory strain *S. cerevisiae* W303 (*MATa, leu2, trp1, ura3, ade2, his3, can1*) [231].

4.2.1.1 Environmental Condition I: Glucose as a Carbon Source

Colonies of *S. cerevisiae* W303 or yAO6 were grown on yeast extract-peptone-dextrose (YPD) agar plates at 30°C for 3 days. For each strain, a single colony was used to inoculate 50 ml of liquid YP media containing 2% glucose (YPD liquid media). The culture was incubated overnight at 30°C with shaking at 225 rpm.

4.2.1.2 Environmental Condition II: Galactose as a Carbon Source

Colonies of *S. cerevisiae* W303 were grown on a yeast extract-peptone-dextrose (YPD) agar plate at 30°C for 3 days. A single colony was used to inoculate 50 ml of liquid YP media containing 2% galactose (YPG liquid media). The culture was incubated overnight at 30°C with shaking at 225 rpm.

4.2.1.3 Environmental Condition III: Response of Mating Type a Yeast to Mating Factor α

Colonies of *S. cerevisiae* W303 or yAO6 were grown on yeast extract-peptone-dextrose (YPD) agar plates at 30°C for 3 days. For each strain, a single colony was used to inoculate 50 ml of liquid YP media containing 2% glucose (YPD liquid media). The culture was incubated overnight at 30°C with shaking at 225 rpm. Fifty microlitres of a 2 μg/μl solution of *S. cerevisiae* mating factor α (yeast pheromone; Sigma Aldrich) was added with mixing to the culture, and incubation continued for 30 minutes.

4.2.1.4 Environmental Condition IV: Response of Yeast to High Osmolarity

Colonies of *S. cerevisiae* W303 were grown on a yeast extract-peptone-dextrose (YPD) agar plate at 30°C for 3 days. A single colony was used to inoculate 50 ml of liquid YP media containing 2% glucose (YPD liquid media). The culture was incubated overnight at 30°C with shaking at 225 rpm. A 5 M NaCl solution was added to bring the final salt concentration of the culture to 0.4 M, and incubation continued for 15 minutes.

4.2.2 RNA preparation from *S. cerevisiae* by phenol/freeze method

All plasticware and laboratory materials used in the preparation or handling of RNA samples were cleaned thoroughly, treated with diethyl pyrocarbonate (DEPC), and autoclaved. All reagents were either prepared using double-distilled and deionized water treated with DEPC, or were DEPC-treated directly.

Cells were harvested by centrifugation, resuspended in 400 µl of AE buffer (50 mM NaAc, 10 mM EDTA, pH 5.2), and transferred to a 1.5 ml Eppendorf tube. Forty microlitres of 10% SDS were added, and the cells were vortexed. One volume (440 µl) of AE-equilibrated phenol was added. Following vortexing and incubation at 65°C for 4 min, the cells were rapidly chilled by immersing the tube in a dry-ice/ethanol bath until phenol crystals were visible within the tube. Centrifugation was performed at room temperature for 2 min, and the aqueous phase transferred to a new tube. An equal volume (500 µl) of P/C/IAA was added and the aqueous suspension was vortexed and centrifuged for 5 min at room temperature. The aqueous phase (~450 µl) was again transferred to a fresh tube, where 1/10 volume (45 µl) of 3M NaAc (pH 5.2) and 2.5 volumes (1240 µl) of ethanol were added. The RNA was allowed to precipitate overnight at -20°C. Finally, the RNA was pelleted by centrifugation, washed with 80% ethanol, allowed to air-dry, and resuspended in 20 µl DEPC-treated ddH_2O. For each of the four environmental conditions, the duplicated RNA samples were pooled, providing a total working volume of 60 µl for each sample. Three microlitres of each sample were removed for RNA quantitation and denaturing agarose gel electrophoresis. The remaining 57 µl of each sample were frozen until required. One microlitre of each total RNA sample was used to determine RNA concentration by

UV spectrophotometry. The concentration of total RNA was approximately 10 µg/µl in each sample.

The quality of yeast total RNA was analyzed by denaturing agarose gel electrophoresis. Five hundred milligrams of agarose were dissolved by boiling in 42 ml DEPC-treated ddH_2O. The boiled solution was cooled to 60°C, and 5 ml 10x MOPS (200 mM MOPS, 50 mM NaAc, 10 mM EDTA, pH 7.0) and 2.6 ml formaldehyde were added. The solution was poured into a plastic gel form and allowed to solidify. The gel was submerged in 1x MOPS in a submarine gel electrophoretic apparatus. Five micrograms of each total RNA sample were volume-adjusted to 5.5 µl with DEPC-treated ddH_2O. Five microlitres of formaldehyde loading buffer (56% formamide, 1x MOPS, 15% glycerol, 15 mM formaldehyde, 4.3 µM bromophenol blue) were added to each sample, which was then vortexed, centrifuged briefly and loaded onto the gel. Electrophoresis was performed for 2 h at 5 V/cm. The gel was carefully removed and placed in a 2 mM EtBr staining solution for 20 minutes. The EtBr-stained gel was visualized by UV transillumination and documented as previously described.

4.2.3 Isolation of mRNA populations from yeast total RNA

mRNA was isolated from each of the yeast total RNA preparations using the MicroPoly(A) oligo-dT cellulose filtration kit (Ambion Inc., Austin TX) with minor variations from the manufacturer's suggested protocol. For each yeast total RNA sample from which mRNA was to be isolated, an aliquot containing 400 µg total RNA was transferred to an RNAse-free 1.5 ml tube. One-tenth volume of 5 M NaCl was added and mixed thoroughly. The volume of the mixture was brought to 1 ml through the addition of 910 µl Binding Buffer. The RNA was heated to 65°C for 5 min, and then immediately chilled on ice for 1 min. A vial of oligo-dT cellulose was added and mixed by inversion for 1 h at room temperature to suspend the resin in the RNA solution.

Centrifugation at 3000x g for 3 min pelleted the oligo-dT resin to which poly(A) RNA was bound, and the supernatant was removed. One millilitre of Binding

Buffer was added as a wash and the resin resuspended by pipetting up and down. Centrifugation and washing were repeated twice using Binding Buffer, then three more times using Wash Buffer. Following the final wash, the resin was resuspended in 400 µl Wash Buffer, and the oligo-dT resin was transferred to a spin column in a 2-ml microcentrifuge tube. After centrifugation at 5000xg at room temperature for 10 sec, 500 µl Wash Buffer were added to the oligo-dT cellulose trapped on the spin column filter pad. A total of three column washes by centrifugation followed. Finally, 100 µl Elution Buffer (prewarmed to 65°C) were added to the oligo-dT resin and the column spun into a new 2-ml tube at 5000xg for 30 sec. A second 100-µl aliquot of Elution Buffer was added and a second elution spin performed, resulting in a total of 200 µl eluted mRNA solution for each sample.

The mRNA was precipitated with 20 µl 5M NH_4OAc, 2 µl glycogen and 500 µl 100% ethanol, and was left to precipitate overnight at -20°C. The precipitated mRNA was recovered by centrifugation at 12 000x g for 30 min at 4°C. The supernatant was removed by aspiration, and the pelleted mRNA was resuspended in 10 µl 0.1M EDTA.

Quantitation of mRNA was performed by the ethidium bromide spot assay. This simple method of mRNA quantitation, useful for evaluating very small volumes of dilute nucleic acid solutions, is capable of detecting less than 5 ng of RNA, and is accurate within an error range of two. RNA concentration standards from which a standard curve could be constructed were prepared from Ambion control mouse lung mRNA (500 ng/µl). Standards prepared from this stock were diluted to 250 ng/µl, 125 ng/µl, 80 ng/µl, 50 ng/µl, 40 ng/µl, and 25 ng/µl. Equal volumes of each standard were mixed with 2 ng/µl EtBr. A control consisting only of 2 ng/µl EtBr was also used. In order to conserve the mRNA samples for use in cDNA synthesis, the volumes of mRNA samples used were as small as was practical. Half a microlitre of each mRNA sample was added to 0.5 µl DEPC-treated ddH_2O, and 1 µl of 2ng/µl EtBr was mixed with the sample. Two-microlitre aliquots of the diluted mRNA standards and of the yeast mRNA samples were spotted onto Saran Wrap placed on a UV transilluminator. The spots were labeled by marking the Saran Wrap adjacent to each spot, and a digital image of the standard curve and the experimental samples was captured. The intensities of each of the standard spots and each of the sample spots were evaluated by plotting signal intensity profiles of the grayscale spot cross-sections using NIH Image v1.62. From the results of this assay, the concentrations of the yeast mRNA sample stocks were determined to be approximately 250 - 400 ng/µl.

Control amplifications from yeast mRNA stocks with the ACT1 and STE2 TF primer sets (TABLE 4.2) were performed to ensure that no yeast genomic DNA was present in the yeast mRNA samples. Twenty nanograms of each mRNA tested was added to a *Taq* DNA polymerase amplification reaction containing 20 pmol of the appropriate TF forward primer and 20 pmol of the matched TF reverse primer. Amplifications proceeded for 30 cycles using standard indexing PCR parameters. Controls were analyzed by agarose gel electrophoresis.

4.2.4 cDNA synthesis: first-iteration protocol

4.2.4.1 First strand cDNA synthesis by RETROscript method using anchored poly(T)$_{16}$-V primers

Early preparations of cDNA populations from yeast mRNA populations employed a first-strand cDNA synthesis protocol based on the RETROscript Reverse Transcription Kit (Ambion) and priming first-strand synthesis with anchored poly(T)$_{16}$-V or anchored poly(T)$_{35}$-V oligonucleotides (where V is a equimolar mix of A, C and G). In addition to the experimental yeast mRNA stocks, first-strand cDNA synthesis was also performed on control mouse lung mRNA (Ambion) and control mouse liver total RNA (Ambion) placed directly into the RETROscript reactions without poly(A) RNA purification. These additional mRNA sources were used as controls and as expendable diagnostic stocks for cDNA quality evaluation.

Each first-strand synthesis reaction contained between 1.5 and 2 µg mRNA, 40 pmol anchored poly(T)$_{16}$-V primer or anchored poly(T)$_{35}$-V primer, and 10 nmol dNTPs in a volume of 16 µl. The reaction contents were mixed, centrifuged briefly, and heated to 80°C for 3 min and placed on ice. Alternate First-strand Buffer (10x:

750 mM KCl, 30 mM $MgCl_2$, 50 mM DTT, 500 mM Tris-HCl, pH 8.3), 10 U placental RNAse inhibitor, and 100 U Moloney-Murine Leukemia Virus Reverse Transcriptase (MMLV RT) were added for a total reaction volume of 20 µl. The reactions were mixed and incubated at 42°C for 1 h. A 10-min incubation at 92°C was used to denature the reverse transcriptase. The reactions were stored at -20°C until required for control amplification or for continuation with second strand cDNA synthesis.

Controls were in general performed as suggested by the manufacturer. Additional positive controls utilized included amplification by the kit's control primer set of the transcript of the murine "housekeeping" gene *rig/S15* from mouse lung first-strand cDNA and from mouse liver first-strand cDNA synthesized from the poly(A) fraction of unpurified total RNA.

4.2.4.2 Second strand cDNA synthesis by modified Klenow method

Early preparations of cDNA populations from yeast first-strand cDNA populations employed a second-strand cDNA synthesis protocol based loosely on that outlined by Sambrook, Fritsch and Maniatis [49]. To the 20-µl volumes of the first-strand cDNA syntheses, 22.5 µl of second-strand synthesis mix (4.25 mM $MgCl_2$, 9 mM $(NH_4)_2SO_4$, 180 µM dNTPs, 60 mM Tris-HCl, pH 7.4), 0.5 µl RNAse H, and 2 µl *E. coli* DNA polymerase I were added for a total reaction volume of 55 µl. The reaction was incubated for 12 h at 16°C. Five microlitres of *E. coli* DNA ligase (10 U/µl; New England BioLabs) were added, as was enough 10x *E. coli* DNA Ligase Buffer (100 mM $MgCl_2$, 100 mM DTT, 260 µM NAD^+, 250 µg/ml BSA, 500 mM Tris-HCl, pH 7.8) to bring the reaction volume to 66.7 µl. The ligation reaction was incubated for 12 h at 16°C, and 3.3 µl 0.5 M EDTA were added to halt the reaction. The cDNA was ethanol-precipitated through the addition of 7 µl 5 M NH_4OAc, 1 µl glycogen and 200 µl cold 95% ethanol, and allowed to precipitate overnight. The tubes containing the yeast cDNA populations were centrifuged to pellet the double-stranded (ds) cDNA, the pellets dried and resuspended in ddH_2O, and quantitated by UV spectrophotometry.

The quality of yeast ds-cDNA populations produced through second-strand cDNA synthesis was evaluated by target fragment amplification from ds-cDNA transcripts using the Test Fragment primer pairs. One hundred nanograms of second-strand cDNA and 20 pmol of each target-specific primer were used in each 50-µl PCR reaction. Standard PCR conditions for the amplification of fragments by *PfuTurbo*™ DNA polymerase were used. Controls were analyzed by agarose gel electrophoresis according to standard protocols for analysis indexed DNA fragments.

4.2.5 Digestion of double-stranded cDNA populations by *Fok*I restriction endonuclease

Digestion of yeast ds-cDNA populations were performed in a manner similar to that described for digestion of pUC19 DNA with *Fok*I restriction endonuclease. Digestions were cleaned of protein and reaction salts by chloroform extraction followed by ethanol precipitation. To establish an approximation of the range of ds-cDNA stock concentrations, quantitation of the ds-cDNA population derived from mouse liver total RNA was performed by UV spectrophotometry.

Completion of *Fok*I digests of yeast ds-cDNA populations was evaluated by target fragment amplification from *Fok*I-digested ds-cDNA transcripts using the TF primer pairs. Twenty nanograms of *Fok*I-digested ds-cDNA and 20 pmol of each target-specific primer were used in each 50-µl PCR reaction. Standard PCR conditions for the amplification of fragments by *PfuTurbo*™ DNA polymerase were used. Controls were analyzed by agarose gel electrophoresis using a standard protocol for indexed DNA fragment analysis.

4.2.6 Amplification of biotinylated indexed cDNA fragments from *Fok*I-digested cDNA populations with BamCC and TF primers following Dynabead extraction

Twenty nanograms of the appropriate *Fok*I-digested yeast ds-cDNA stock was added to a single-indexer ligation reaction containing 500 fmol of biotinylated nonphosphorylated BamCC indexer mix specific for the 3'-end *Fok*I fragment of either ACT1, FUS1, STE2 or STE12. Ligation with 4 U of T4 DNA ligase was performed for 14 hrs at 16°C.

Streptavidin-coated paramagnetic beads (Dynabeads M-280) (Dynal AS, Oslo, Norway) were rinsed according to the supplier's instructions using a Dynal Magnetic Particle Concentrator (MPC). The rinsed beads were resuspended and 16 µl removed to a 200-µl PCR tube. The tube was placed for 3 minutes in a home-built magnetic particle concentrator rack using 1300 milliTesla NbFeB magnets (2 mm x 5 mm diam.) fused into the side of eight positions of a 96-well tube holder. The supernatant was pipetted off the beads, which were then suspended in 160 µl 2 M NaCl. Twenty microlitres of the washed Dynabead suspension were added to the 20-µl biotinylated-indexer ligations, resulting in a final hybridization concentration of 1 M NaCl. The biotinylated indexing ligation reactions were incubated with the Dynabead suspension at 37°C for 1 h with rotation. Following incubation, the Dynabeads were collected using the home-built MPC and the supernatant was removed. The beads were washed twice with 200 µl of 2 M NaCl prior to a final wash with ddH_2O and addition of the beads to an amplification reaction.

Each 50-µl PCR reaction contained 20 pmol BamCC indexing primer and 20 pmol of the appropriate target-specific primer. Standard PCR conditions for the amplification of fragments by *Taq* DNA polymerase were utilized. Amplified reactions were analyzed by agarose gel electrophoresis.

4.2.7 Modified protocol for ds-cDNA synthesis from *S. cerevisiae* mRNA populations

4.2.7.1 cDNA synthesis using GCRichPoly(T)$_{16}$-V primers and SuperScript™ System

Later preparations of cDNA populations from yeast mRNA populations employed a first-strand cDNA synthesis protocol based on the SuperScript cDNA Synthesis Kit (Gibco-BRL Life Technologies). In these later preparations, first-strand synthesis was primed with the anchored poly(T) variant oligonucleotide mix GCRichPoly(T)$_{16}$-V.

First-strand cDNA synthesis using GCRichPolyT$_{16}$-V primers

For each cDNA population to be synthesized, 7 µl (2-3 µg) of the appropriate mRNA stock was added to 1 µl of GCRichPoly(T)$_{16}$-V (200 pmol total oligo) in a sterile 1.5 ml microcentrifuge tube. The mixture was heated to 70°C for 10 min, and rapidly chilled on ice. The contents of the tube were collected by a brief centrifugation step. Four microlitres of 5x First Strand Buffer [250 mM Tris-HCl (pH 8.3), 15 mM $MgCl_2$, 375 mM KCl], 2 µl 0.1 M DTT, and 1 µl 10 mM dNTP mix were added, providing a total volume of 15 µl. The contents were mixed by vortexing, collected by centrifugation and incubated at 37°C for 2 min to allow for temperature equilibration. Five microlitres (1000 U) of SuperScript II reverse transcriptase (a cloned variant of M-MLV RT modified to remove RNAse H activity) were added, mixed gently by pipetting, and the reaction incubated for 1 hr at 37°C. First-strand cDNA synthesis was halted by placing the reaction tubes on ice.

Second-strand cDNA synthesis

With the reaction tubes on ice, second-strand synthesis components were added in the following order: 91 µl DEPC-treated ddH_2O; 30 µl 5x Second Strand Buffer [100 mM Tris-HCl (pH 6.9), 450 mM KCl, 23 mM $MgCl_2$, 0.75 mM β-NAD$^+$, 50 mM $(NH_4)_2SO_4$]; 3 µl 10mM dNTP mix; 1 µl *E. coli* DNA ligase (10 U/µl); 4 µl *E. coli* DNA polymerase I (10 U/µl); and 1 µl *E. coli* RNAse H (2 U/µl). The assembled second-strand synthesis reaction (total volume 150 µl) was mixed by gentle vortexing, centrifuged briefly to collect the reaction contents, and incubated in a electronically-regulated cold water bath for 2 hours at 16°C. Two microlitres of T4 DNA polymerase (5 U/µl) were added, the reaction was incubated for a further 5 min at 16°C, and synthesis was halted by placing the reaction on ice and adding 10 µl of 0.5 M EDTA.

To extract the double-stranded cDNA from the various reaction components, 150 µl of P/C/IAA (25:24:1) was added to the halted synthesis reaction. Thorough vortexing was followed by a 5-min centrifugation at 14 000 rpm to separate the organic and aqueous phases, and 140 µl of the aqueous layer was transferred to a new 1.5-ml microcentrifuge tube. One hundred microlitres of 5 M NH_4OAc were added, followed by 500 µl of chilled (-20°C) absolute ethanol. The mixture was vortexed thoroughly and immediately centrifuged for 20 min at 14 000 rpm at room temperature. The supernatant was pipetted off, and the cDNA pellet was washed with chilled 70% ethanol. A second centrifugation step followed, and the supernatant was again removed. The cDNA pellet was air-dried at 37°C for 10 min, then resuspended in 10 µl 0.1 M EDTA. The concentration of each cDNA population was determined by UV spectrophotometry. Each ds-cDNA population was stored at -20°C until required.

Second-strand synthesis controls

The quality of yeast ds-cDNA populations produced through the modified cDNA synthesis approach was evaluated by target fragment amplification from ds-cDNA transcripts using the TF primer pairs in the manner previously described.

4.2.7.2 Fok*I digestion of ds-cDNA populations primed with GCRichPoly(T)$_{16}$-V:*

Digestion of yeast ds-cDNA populations produced through the modified cDNA synthesis approach with *Fok*I restriction endonuclease was performed as previously described.

4.2.8 Optimized ligation and amplification conditions for 3'-end cDNA indexing

A typical 3'-end cDNA indexing ligation reaction contained 1 ng of *Fok*I-digested yeast ds-cDNA, 500 fmol of the appropriate biotinylated nonphosphorylated BamCC indexer mix and 10 U *Taq* DNA ligase (40 U/µl) in a 15-µl reaction volume of *Taq* DNA ligase buffer [20 mM Tris-HCl (pH 7.6 @ 25°C), 25 mM KAc, 10 mM DTT, 1 mM NAD^+, 0.1% Triton X-100]. Whenever practical, assembly of cDNA indexing ligations, Dynabead extractions and amplifications were performed in parallel using master reaction mixes, multi-channel pipettors and 96-well-plate-compatible formats. cDNA ligations were incubated at 16°C for 2 h.

Twelve microlitres of unrinsed Dynabead suspension was rinsed as suggested by the manufacturer and resuspended in 15 µl of 2 M NaCl. This volume of washed streptavidin-coated beads was then added to a completed cDNA indexing ligation containing biotinylated indexers and, therefore, biotinylated cDNA fragments. The samples to be extracted were incubated at 37°C for 1 h with tumbling rotation. The beads were washed with 100 µl 2xSSC/50% formamide, rinsed with 100 µl ddH_2O, and the liquid removed using a single-channel pipettor and the home-built MPC.

A 18-µl volume of Platinum™ *Taq* DNA polymerase reaction mix was prepared with 40 pmol fluorescently-labeled BamCC indexing primer; 60 pmol [total oligo] GCrichPolyT_{16}-V (20 pmol/anchored end); 2 µl 10x PCR Buffer; 1.5 µl of 10 mM dNTPs; 1.5 µl of 50 mM $MgCl_2$; and 1.25 U Platinum™ *Taq* DNA polymerase (2.5 U/µl). BamCC indexing primers were labeled with one of the fluorescent dyes FAM, JOE, or ROX. In instances in which the amplified fragments were to be analyzed by automated DNA sequencing instrumentation and compared with indexed cDNA amplifications representative of yeast gene expression under varying environmental growth conditions, a different colour of dye-labeled primer was used for each yeast cDNA source. This PCR reaction mix was added onto the magnetically-isolated Dynabeads to which the biotinylated cDNA fragments were bound, and mixed thoroughly by pipetting up and down. 3'-end cDNA targets bearing a biotinylated BamCC indexer on one end and a GCRichPoly(A) region at the other were amplified for 35 cycles using the following cycling parameters: (30 sec at 92°C; 30 sec at 55°C; 1 min at 72°C) for 2 cycles; (30 sec at 92°C; 30 sec at 59°C; 1 min at 72°C) for 33 cycles; 10 min at 72°C; hold at 4°C.

To prepare samples for analysis by agarose gel electrophoresis, 4 µl of loading dye was added directly to the 20-µl PCR reaction and mixed thoroughly. Twelve microlitres of each dyed reaction was loaded onto a 2% agarose gel and electrophoresed for 120 min at 8 V/cm. Appropriate DNA size standards were included on the gel. Visualization of electrophoresis results was performed by UV transillumination as previously described, and the data were recorded by digital image capture.

Alternately, fluorescently-labeled cDNA indexing samples were analyzed using automated DNA sequencing instrumentation. Prior to analysis, 3'-end cDNA amplifications indexed with identical nonphosphorylated BamCC indexer mixes but originating from different yeast cDNA populations were pooled. For the case in which levels of gene expression from three different growth conditions were to be compared, three indexing PCR reactions targeting the same cohesive-end sequence but originating from different yeast cDNA populations and thus bearing three differently-coloured fluorescent tags were pooled for a total volume of 60 µl. The pooled samples were purified with the Concert™ Rapid PCR Purification System and concentrated into 5 to 10 µl with Microcon YM-30 spin columns. A TAMRA-labeled DNA size standard (LargeFrag DNA size standard) was generated by amplifying 666-bp, 826-bp and 983-bp fragments from a duck hepatitis B virus (D-HBV) size standard set with a TAMRA-labeled common forward primer [232], and the reactions were pooled and concentrated in a similar manner to the indexed cDNA fragments. Two microlitres of GeneScan-500 [TAMRA] DNA size standard (Applied Biosystems) and 2 µl of the LargeFrag DNA size standard were added to each concentrated reaction and mixed thoroughly. The standard-doped indexed cDNA reactions were stored at -20°C in the dark until analyzed.

4.2.9 3'-end cDNA indexing data acquisition and analysis by automated DNA sequencing instrumentation

Data acquisition for fluorescently-labeled 3'-end cDNA fragment sets doped with TAMRA-labeled DNA standards was performed on an ABI 377 DNA Sequencer (Applied Biosystems) according to the manufacturer's instructions for RFLP genetic analysis. Experimental data sets were analyzed and manipulated using ABI GeneScan v3.1 genetic analysis software for automated sequencers.

4.3 RESULTS

4.3.1 Selection of Test Fragment (TF) target sequences and design of TF primer pairs

Four gene sequences were selected to serve as templates for controls in multiple aspects of cDNA indexing protocol development: *ACT1*, *FUS1*, *STE2* and *STE12*.. The *ACT1* gene product, actin, is a major cytoskeletal protein with roles in endocytosis, cell polarization and nuclear migration [233]. The gene is consititutively active, with roughly 60 copies of the *ACT1* mRNA present per cell during logarithmic growth in glucose medium [214]. Cells grown under these conditions contain approximately 60 000 copies of the actin molecule.

FUS1 is expressed only in haploid cells [234] and is essential for cell fusion during mating [235]. The *FUS1* promoter sequence contains four pheromone-responsive elements (PREs) [236] which are bound by the pheromone-induced transcription factor Ste12p [237]. A very small number of *FUS1* mRNA molecules with half-lives of less than 3 minutes are expressed in MATa cells grown in the absence of mating factor [238]. A rapid 100-fold increase in *FUS1* transcription is induced within fifteen or twenty minutes of the addition of α mating factor (pheromone) to MATa yeast cultures [239, 240].

The G protein-coupled receptor for α mating factor is the protein product of the constitutively active *STE2* gene [241]. This gene product is essential for the mating of MATa cells with MATα haploids, as it is required for shmoo formation, agglutination and cell cycle arrest in pheromone-treated MATa cells [242]. The abundance of *STE2* mRNA, which has a 4- to 5-minute half-life [238], is roughly 2 to 10 copies per cell during exponential growth in glucose medium [195] and increases slightly in response to pheromone.

Ste12p is the transcription factor that binds to PREs for the regulation of genes required for mating [234] and is also involved in gene regulation for pseudohyphal formation and filamentous growth [243]. It is required for the pheromone-induced transcriptional activation of almost thirty genes [237]. *STE12* is constitutively expressed at low levels, and its transcription is moderately induced by an upstream MAP kinase cascade pathway in response to the binding of α mating factor (pheromone) to the Ste2p G protein-coupled receptor [234, 237].

This combination of four gene sequences represented a target set exhibiting a range of basal transcription levels and expression patterns. Primer pairs designed to amplify Test Fragments from cDNA transcripts of the *ACT1*, *FUS1*, *STE2* and *STE12* mRNAs were useful in controlling for key steps of cDNA synthesis. When employed in combination with indexers, these primer pairs were valuable tools in the modification and evaluation of 3'-end cDNA indexing protocols for global gene expression studies of *S. cerevisiae*. The Web Primer tool of the Saccharomyces Genome Database (http://genome-www.stanford.edu/Saccharomyces) was used to determine a matched pair of optimal primer sequences amplifying a fragment of at least 500 bp from the 3'-end of the cDNA transcript of each gene of interest (TABLE 4.1). The Test Fragment sequences anticipated from amplification by the transcript-specific primer pairs were scanned for *Fok*I recognition sites using Sequencher v3.0 software. The lengths and cohesive end sequences of 5' and 3' restriction fragments expected from *Fok*I cleavage of the Test Fragments were determined (TABLE 4.2 and FIGURE 4.1).

4.3.2 Use of Test Fragments as controls

4.3.2.1 Amplification of Test Fragments as controls for second-strand cDNA synthesis quality

The quality of yeast ds-cDNA populations was evaluated by target amplification from ds-cDNA transcripts using the TF primer pairs. The amplification of Test Fragments from ds-cDNA synthesized using anchored poly$(T)_{16}$-V primers (**Section 4.2.4**) from mRNA isolated from pheromone-treated cultures of *S. cerevisiae* strain W303 is shown in FIGURE 4.3A. Each TF primer pair successfully amplified its target from cDNA transcripts present in the Pheromone ds-cDNA population.

TABLE 4.1: Sequences of Test Fragment primer pairs.

Test Fragment	Primers			
	Primer Name	Sequence	Length (nt)	T_m (ºC)
ACT1	ACT1 Forward	ATCCAAGCCGTTTTGTCCTTGTA	23	58.6
	ACT1 Reverse	ATGGACCACTTTCGTCGTATTC	22	56.1
FUS1	FUS1 Forward	CACGCCAGATTCACAAATCA	20	54.6
	FUS1 Reverse	CAGTCGTATTCTTGGAGACAGTCA	24	57.6
STE2	STE2 Forward	CCACAATTTTACTTGCATCCTC	22	53.5
	STE2 Reverse	TACATGTCGACGGGTTCAACTT	22	57.8
STE12	STE12 Forward	GGATTTTGATGAATCTCGGC	20	52.6
	STE12 Reverse	GGCATCTGGAAGGTTTTTATCGG	23	57.6

TABLE 4.2: Description of Test Fragments.

Target Locus	Test Fragment	5'-*Fok*I Fragment		3'-*Fok*I Fragment		
	Length (bp)	Indexer Sequence	Length	Indexer Sequence	Length (Primer)	Length (Poly$_{(A)}$ Tail)
ACT1	697 bp	AGAN	394 bp	TTCN	351 bp	401 bp
FUS1	534 bp	AGAN	449 bp	CTCN	133 bp	158 bp
STE2	610 bp	ATCN	165 bp	CTGN	324 bp	415 bp
STE12	816 bp	TTCN	266 bp	TATN	169 bp	200 bp

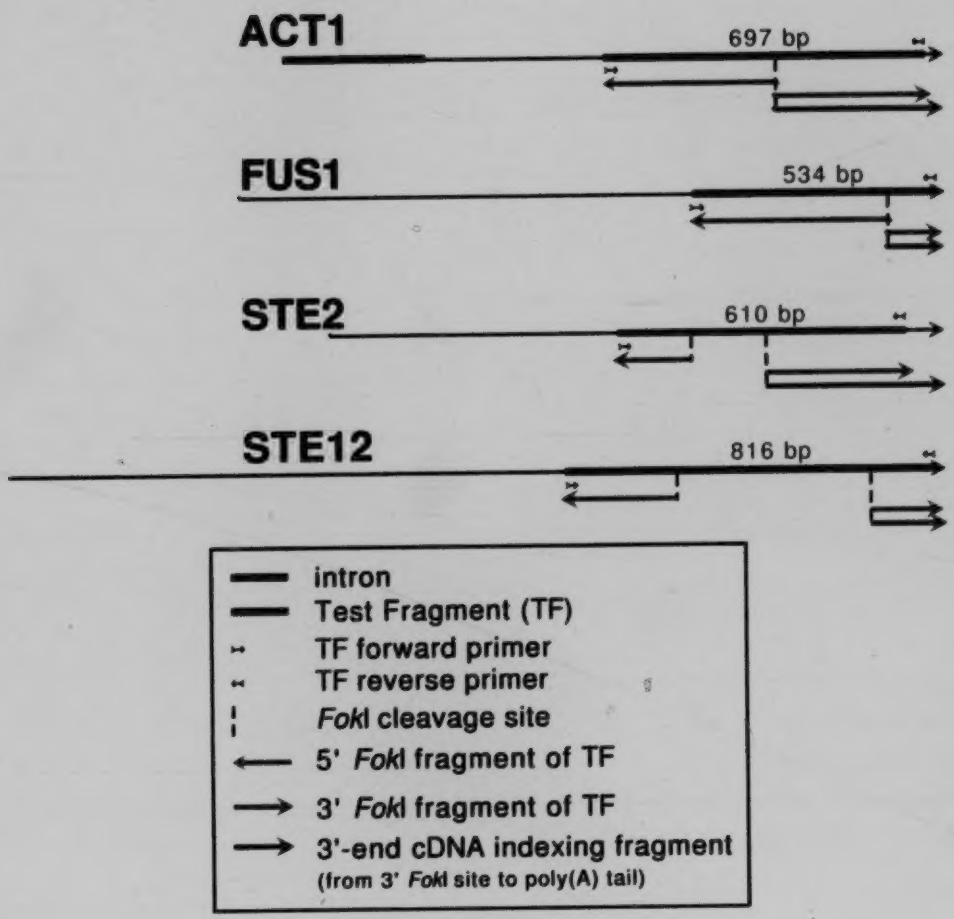

ACT1
697 bp
FUS1
534 bp
STE2
610 bp
STE12
816 bp
intron
Test Fragment (TF)
TF forward primer
TF reverse primer
FokI cleavage site
5' *FokI* fragment of TF
3' *FokI* fragment of TF
3'-end cDNA indexing fragment
(from 3' *FokI* site to poly(A) tail)

FIGURE 4.2 Design of Test Fragment target sequences. See text and Table 4.2 for details.

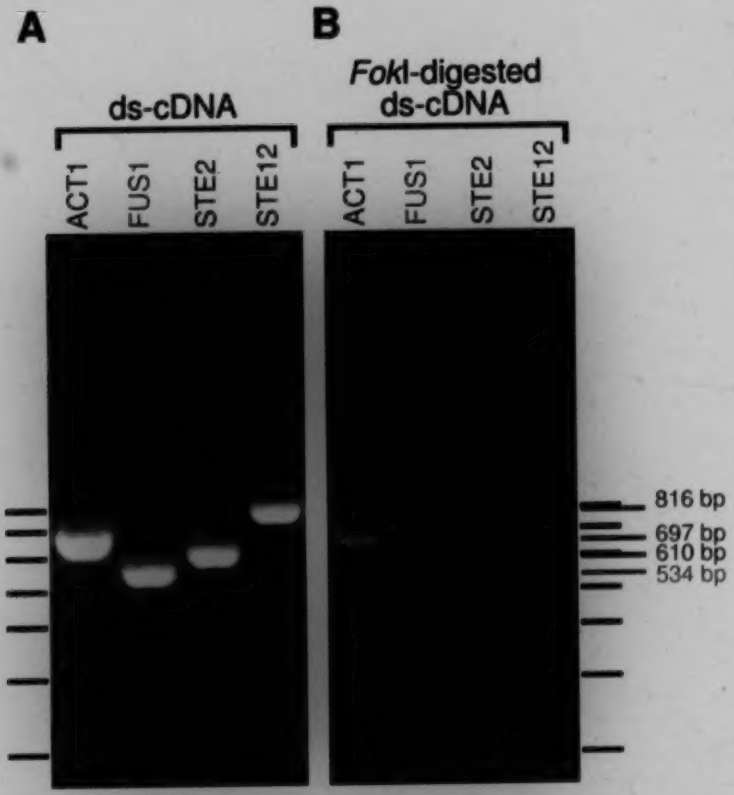


FIGURE 4.3: Use of Test Fragments as controls for second-strand cDNA synthesis quality and for ds-cDNA *FokI* digest quality.
A) Second-strand cDNA synthesis quality was evaluated by Test Fragment amplification using transcript-specific primer pairs. First-strand cDNA synthesis from mRNA isolated from pheromone-treated cultures of *S. cerevisiae* strain W303 was initiated from anchored poly(T)$_{16}$-V primers. Second-strand cDNA synthesis was performed as described in the text. PCR reactions containing 20 ng of Pheromone ds-cDNA and 20 pmol of each of two transcript-specific TF primers were assembled in a total volume of 50 µl. Standard conditions previously described for the amplification of indexed bacterial DNA fragments by *PfuTurbo*™ DNA polymerase were used. Agarose gel electrophoresis was performed as previously outlined.
B) Quality of *FokI* digests of ds-cDNA populations was evaluated by Test Fragment amplification. Double-stranded cDNA populations synthesized as described above were digested with *FokI* restriction endonuclease. PCR reactions containing 20 ng of *FokI*-digested Pheromone ds-cDNA and 20 pmol of each primer in a TF primer pair were assembled in a total volume of 50 µl. PCR and agarose gel electrophoresis analysis of amplification products were performed as described previously.

Similar controls were performed for each ds-cDNA population synthesized (data not shown).

4.3.2.2 Amplification of Test Fragments as control for ds-cDNA Fok*I digest quality*

The quality of *Fok*I restriction digests of ds-cDNA populations derived from yeast cultures subjected to various environmental stimuli was evaluated using TF primer pairs. Amplification of each Test Fragment from undigested ds-cDNA populations was attempted. *Fok*I digestion of the Pheromone ds-cDNA population was sufficiently complete to prevent amplification of the *FUS1*, *STE2* and *STE12* TFs from undigested *FUS1*, *STE2* or *STE12* ds-cDNA templates remaining following incubation with *Fok*I endonuclease (FIGURE 4.3B). Amplification of the *ACT1* Test Fragment by the *ACT1* TF primer pair was significantly reduced, but not completely eliminated, by *Fok*I digestion. The high level of constitutive expression of the *ACT1* mRNA transcript in *S. cerevisiae* cells resulted in a high copy number of the *ACT1* ds-cDNA transcript in the complete ds-cDNA population. Even under highly efficient *Fok*I digestion conditions only a very small percentage of the numerous *ACT1* ds-cDNA transcripts needed to avoid digestion in order to permit amplification of the *ACT1* TF. The use of longer incubation times or higher concentrations of restriction endonuclease failed to completely inhibit *ACT1* TF amplification, and were found to contribute to increased non-specific degradative activity of the enzyme on the ds-cDNA population (data not shown). Insofar as a high percentage of each transcript was accurately and specifically digested, permitting efficient downstream indexing and amplification of targeted transcript restriction fragments, the restriction digestion conditions employed were suitable for 3'-end cDNA indexing purposes. Similar results were obtained for each *Fok*I-digested cDNA population analyzed (data not shown).

4.3.3 Use of *Fok*I-digested Test Fragments as artificial target fragments for DNA indexing

In order to determine the minimum amount of a particular *Fok*I-digested cDNA transcript necessary for successful targeting and amplification by 3'-end cDNA indexing, artificial target fragments were constructed from Test Fragment amplicons. Following preparative amplification, TF amplicons were digested by *Fok*I endonuclease and diluted to form amplicon stocks of known concentration. An initial set of indexing reactions were performed to demonstrate that 5'- and 3'-terminal restriction fragments derived from *Fok*I-digested TF amplicons could be targeted by indexer ligation and amplified using an indexing primer and the appropriate directional TF primer. A second set of indexing reactions was then performed on serially-diluted stocks of *Fok*I-digested TF amplicons to determine the minimum number of digested transcript copies required for ligation of targeting indexers and amplification of the indexed amplicons.

4.3.3.1 Preparative amplification and Fok*I digestion of Test Fragment amplicons*

In order to generate artificial indexing target fragments, Test Fragments were amplified in large preparative PCR reactions and digested by *Fok*I endonuclease. For each of the four TF primer pairs, four parallel amplifications were performed using undigested Pheromone ds-cDNA populations as a source of fragment templates. PCR conditions were identical to those used for amplification of second-strand cDNA synthesis controls. Following PCR, the parallel amplifications of each Test Fragment were pooled, and a small aliquot analyzed for purity by agarose gel electrophoresis (data not shown). The pooled Test Fragment amplifications were purified of enzyme and salts and digested by *Fok*I endonuclease. Digestion was performed as described for *Fok*I digestion of pUC19 DNA. Each of the *Fok*I digests of the four amplicons ACT1-TF, FUS1-TF, STE2-TF and STE12-TF were serially diluted to provide amplicons stocks with concentrations of 1 ng/µl, 100 pg/µl, 10 pg/µl, 1 pg/µl, 100 fg/µl, 10 fg/µl, 1 fg/µl and 100 ag/µl.

4.3.3.2 Initial indexing and amplification of Fok*I-digested TF amplicons*

A preliminary effort to index and amplify terminal restriction fragments of *Fok*I-digested TF amplicons was performed using a high concentration of target DNA. This was done to ensure that such amplification was possible and to validate this approach to determining transcript copy number requirements for 3'-end cDNA indexing. Single-indexer-mix ligation reactions were prepared using the 1 ng/µl stock of the relevant *Fok*I-digested TF amplicon and the BamCC NoP indexer mix containing the cohesive end sequence complementary to that of either the 5' or the 3' restriction fragment of the TF amplicon (TABLE 4.2 and FIGURE 4.1). Amplification of the indexed fragment was performed using the BamCC indexing primer and the appropriate directional TF primer (FIGURE 4.4).

For example, as represented in lane A of FIGURE 4.4, the indexing ligation targeting the 5' restriction fragment of the *Fok*I-digested ACT1-TF amplicon (or *ACT1-5'*) contained 2 ng of *Fok*I-digested ACT1-TF DNA and 50 fmol/µl/indexer (200 pmol total indexer) of the NoP indexer OH-AGANxBamCC. This indexer mix included the BamCC indexer with cohesive end sequence AGAA, complementary to the TCTT cohesive end sequence present on the ACT1-5' restriction fragment. Ligation of this indexer to ACT1-5' imported a BamCC primer binding sequence onto one end of the target fragment, which carried the complementary sequence for the ACT1 Forward primer on the other end. Amplification of the 394-bp ACT1-5' target fragment was accomplished successfully.

Amplification of the other artificial target fragments was also successful, although FUS1-5' was amplified only weakly by this approach (lane C of FIGURE 4.4). The indexing and amplification of terminal restriction fragments of *Fok*I-digested TF amplicons at high concentrations of target DNA suggested the possibility of indexing and amplifying these fragments at lower target DNA concentrations.

4.3.3.3 *Evaluation of minimum template concentrations using* Fok*I-digested TF amplicons as models for 3'-end cDNA fragments*

The minimum number of copies of a particular *Fok*I-digested transcript required to enable indexer ligation and target amplification was an important parameter determining the utility of 3'-end cDNA indexing as a method for global gene expression analysis in yeast. The four sets of serially-diluted *Fok*I-digested TF amplicons were useful in estimating the order of magnitude of this parameter. Seven

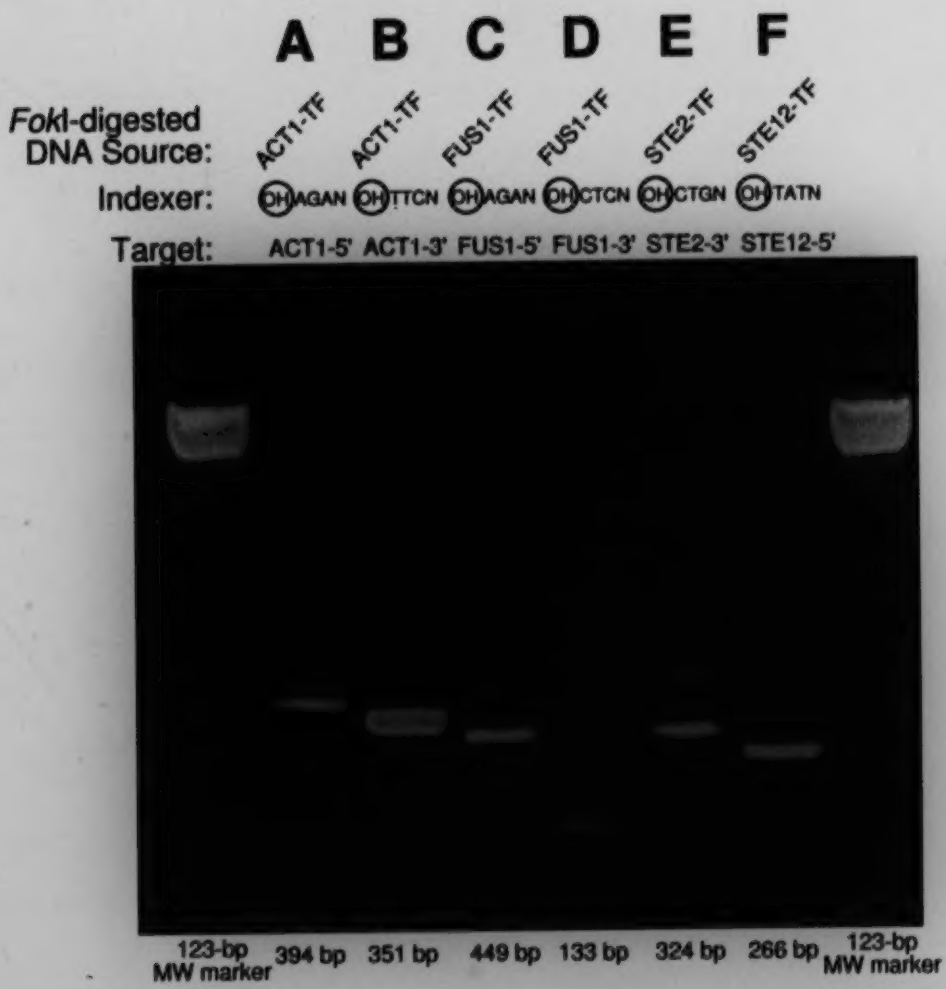

A B C D E F
FokI-digested DNA Source:
ACT1-TF
ACT1-TF
FUS1-TF
FUS1-TF
STE2-TF
STE12-TF
Indexer:
OH AGAN
OH TTCN
OH AGAN
OH CTCN
OH CTGN
OH TATN
Target:
ACT1-5'
ACT1-3'
FUS1-5'
FUS1-3'
STE2-3'
STE12-5'
123-bp MW marker
394 bp
351 bp
449 bp
133 bp
324 bp
266 bp
123-bp MW marker

FIGURE 4.4: Indexing and amplification of *FokI*-digested TF amplicons. Artificial indexing targets were constructed by *FokI* digestion of amplified TF fragments. Single-indexer-mix ligations were assembled with 2 ng of the appropriate *FokI*-digested TF amplicon stock and the BamCC NoP indexer mix specific for either the 5' or the 3' TF restriction fragment. Ligation was performed in a manner previously described for pUC19 indexing ligations. For each indexed TF ligation, a standard PCR reaction was prepared containing BamCC primer and the appropriate directional TF primer. Amplification proceeded using standard indexing PCR conditions employing *PfuTurbo*™ DNA polymerase. Amplified products were analyzed by agarose gel electrophoresis as previously described.

sets of 6 ligation reactions each were assembled, similar in format to the single set of 6 ligations employed in FIGURE 4.4 but containing either 100 pg, 10 pg, 1 pg, 100 fg, 10 fg, 1 fg or 100 ag of the relevant *Fok*I-digested TF amplicon DNA. Ligation, PCR and agarose gel electrophoresis were performed as before (FIGURE 4.5).

Amplification of ACT1-5' (in lane A of FIGURE 4.5), ACT1-3' (lane B), FUS1-3' (lane C), STE2-3' (lane E) and STE12-5' (F) was observed from ligations containing from 100 pg to as little as 10 fg of the relevant target DNA transcript. STE2-3' failed to amplify from a ligation containing 100 pg of target DNA; however, this fragment was efficiently amplified from ligation containing 10 pg to 10 fg of target DNA, indicating that PCR failure in the instance of the 100-pg ligation was artifactual. ACT1-5' was weakly amplified to levels only faintly visible on agarose gel from ligations containing 100 fg or less of target DNA. The FUS1-5' target (lane C) was not efficiently amplified under these conditions, and was not present at detectable levels for target DNA concentrations below 1 pg in ligation. Amplification of indexed *Fok*I-digested TF amplicon fragments to detectable levels was not demonstrated for ligations containing 1 fg or 100 ag of target DNA (data not shown). The excellent reproducibility of DNA indexing was demonstrated across a wide range of target concentrations.

Even for target DNA concentrations as low as 500 ag/µl in ligation (10 fg target DNA in a 20-µl reaction volume), four of the six *Fok*I-digested TF amplicon fragments targeted were efficiently indexed and amplified. Interestingly, each of the 3' terminal restriction fragments targeted in this experiment were efficiently amplified under these conditions. These data permitted rough estimation of the template copy number required for efficient ligation of complementary indexers and subsequent amplification of target fragments using the BamCC indexing primer to primer strand elongation from one end of the target molecule.

Consideration of the case of a hypothetical 3' terminal restriction fragment 400 bp in length when indexed, roughly the size of ACT1-3', was instructive for this estimation. As each molecule of the DNA fragment would have an approximate molecular mass of 260 000, one picogram of this fragment would contain roughly two

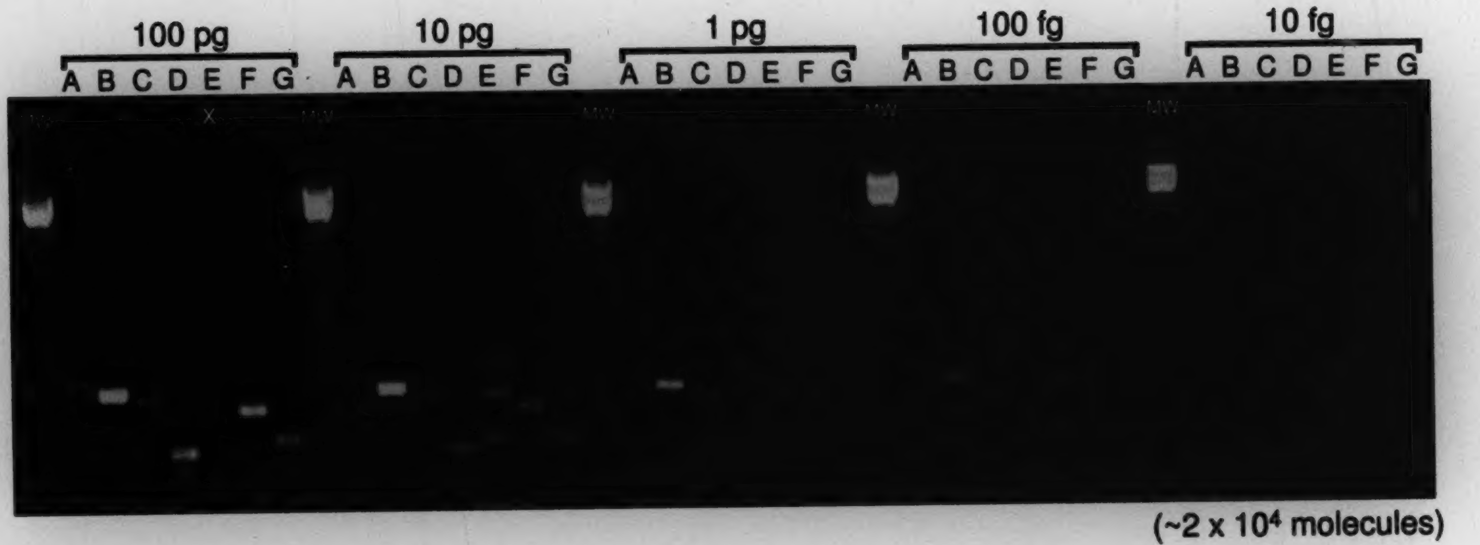


FIGURE 4.5: Estimation of minimum template concentration requirements using *FokI*-digested TF amplicons. A set of serial dilutions (100 pg/µl, 10 pg/µl, 1 pg/µl, 100 fg/µl and 10 fg/µl; results for 1 fg/µl and 100 ag/µl not shown) was prepared from stocks of *FokI*-digested ACT1-TF, FUS1-TF, STE2-TF and STE12-TF amplicons. Single-indexer-mix ligation reactions were prepared using 1 µl of the appropriate *FokI*-digested TF amplicon dilution and the appropriate indexer. For each target DNA concentration, the set of lanes A through F correspond to the same combination of target DNA source, indexer identity and target fragment size as those represented by lanes A through F of FIGURE 4.4. Lane G contains the indexed amplicon corresponding to the 169-bp STE12-3' fragment targeted by OH-TATNxBamCC indexer mix. Ligation and PCR were performed as previously described. Amplification products from each PCR reaction were analyzed by standard agarose gel electrophoresis. Lanes marked "MW" contain 123-bp ladder molecular weight markers.

million molecules. Ten femtograms would therefore represent about twenty thousand template molecules available for targeting by indexer ligation.

It has been estimated that isolations of total cellular mRNA from culture preparations of two million eukaryotic cells should yield 10 μg of total RNA, 3% of which is mRNA [244]. Analysis of abundance distribution of mRNA species within cellular mRNA populations has indicated that a moderately abundant mRNA species may represent about 0.1% of the total mRNA population in a eukaryotic cell [191, 195]. In other words, 10 ng of yeast mRNA would contain 10 pg of a particular mRNA transcript present at a moderately high level of expression.

Of greater significance to the development of applications for the study of global gene expression is the abundance of rare transcripts within total cellular mRNA populations. The vast majority of genes in a eukaryotic genome are expressed at low levels (1-20 mRNA transcript copies per cell), or not at all, except under select growth and environmental conditions [186, 195]. Even under these conditions their transcription may only be transiently upregulated [187, 191]. A specific rare mRNA species may represent 0.003% or less of the total cellular mRNA population of *S. cerevisiae* cells [186, 195]. This corresponds to a range of 3 pg to as little as 30 fg of a 10-ng aliquot of total cellular mRNA.

In the experiment outlined in FIGURE 4.5, specific indexers were ligated to target fragments bearing complementary cohesive end sequences with sufficient efficiency to allow amplification of indexed fragments from ligations containing roughly twenty thousand unindexed target fragment molecules. In ligations containing 10-ng aliquots of *Fok*I-digested yeast ds-cDNA, a particular transcript present at 20 000 copies per ligation represents only one-millionth of the total target cDNA present. Because 10 ng of ds-cDNA represents the total cellular mRNA content of roughly 30 000 yeast cells, 10 fg of a particular cDNA species represents an average expression level approaching one mRNA transcript copy per cell. Coarse estimation of this nature, in combination with the data presented in FIGURE 4.5, suggested the potential sensitivity to rare transcripts which might be achieved by 3'-end cDNA indexing.

The foregoing coarse estimation of rare-transcript sensitivity for cDNA indexing depends on two assumptions. First, it assumes that cDNA synthesis from yeast total cellular mRNA populations results in a double-stranded cDNA population that accurately reflects the relative transcript abundances present in the initial mRNA population. Second, it assumes that the excellent sensitivity demonstrated in the comparatively low complexity of indexing ligations and amplifications of *Fok*I-digested purified TF amplicons reflects the sensitivity available to cDNA indexing when *Fok*I-digested ds-cDNA populations of much higher complexity are targeted. This assumption, and concerns for its validity, were the focus of subsequent cDNA indexing protocol development (**Section 4.3.5**).

4.3.4 Amplification of indexed *Fok*I-digested TF restriction fragments from *Fok*I-digested yeast cDNA populations using BamCC and TF primers

4.3.4.1 Amplification of indexed Fok*I-digested TF restriction fragments directly from* Fok*I-digested yeast cDNA populations*

Although the potential for cDNA indexing to exhibit high sensitivity for detection of rare transcripts had been demonstrated for artificial indexing targets in a DNA population of low complexity, targeting of specific fragments by complementary indexer ligation in a complex *Fok*I-digested cDNA population had not yet been performed. A set of ligations were assembled that targeted several *Fok*I-digested TF terminal restriction fragments from a *Fok*I-digested cDNA population derived from cultures of *S. cerevisiae* strain W303 grown in glucose medium in the absence of pheromone. This cDNA population was selected because of the range of relative abundances manifested by the *ACT1*, *FUS1*, *STE2* and *STE12* transcripts in mRNA populations isolated from this strain grown under these conditions, and the resultant range of abundances expected for the derivative ds-cDNA transcripts.

Five hundred femtomoles (per indexer) of the appropriate NoP indexer mix were ligated to 20 ng of *Fok*I-digested cDNA for 14 hrs at 16°C using 40 U of T4 DNA ligase. These extremely forcing conditions for ligation were employed in order to ensure successful ligation of indexers to targets in this initial attempt to index yeast cDNA restriction fragments directly from a complex population. Amplification of indexed TF terminal restriction fragments using BamCC and the appropriate TF directional primer followed. Thirty-five cycles of PCR were performed using standard conditions for fragment amplification by *PfuTurbo*™ DNA polymerase. Analysis of PCR reaction products revealed the efficient indexing and amplification of the ACT1-3', STE2-3', STE12-5' and STE12-3' targets (FIGURE 4.6A). As *FUS1* transcription is not induced in the absence of pheromone, insufficient levels of *Fok*I-digested *FUS1* cDNA terminal restriction fragments were present in ligation to permit amplification of the FUS1-3' target fragment to detectable levels (data not shown).

In this experiment, specific *Fok*I-digested cDNA transcripts were ligated to indexers bearing cohesive end sequences complementary to those of the target fragments. The indexed targets were then amplified using a common indexing primer (complementary to the indexing primer binding sequence imported onto the *Fok*I-digested end of the target by indexer ligation) and a specific primer complementary to a particular sequence located near the opposite end of the indexed restriction fragment. The successful amplification of target fragments demonstrated the efficient attachment of indexers to target fragments using this approach, an important criterion for the application of 3'-end cDNA indexing. The concomitant amplification of several spurious PCR products in several reactions suggested that reduction of PCR complexity might be required to improve amplification specificity. A method by which to reduce cDNA population complexity between ligation and PCR was sought.

4.3.4.2 Amplification of biotinylated indexed cDNA fragments from Fok*I-digested cDNA populations with BamCC and TF primers following purification by streptavidin-coated paramagnetic beads.*

The use of streptavidin-coated paramagnetic beads was investigated as an attractive method to provide improved amplification specificity through the reduction of cDNA population complexity following ligation. In this approach, biotinylated indexers were ligated to target fragments bearing the appropriate cohesive end

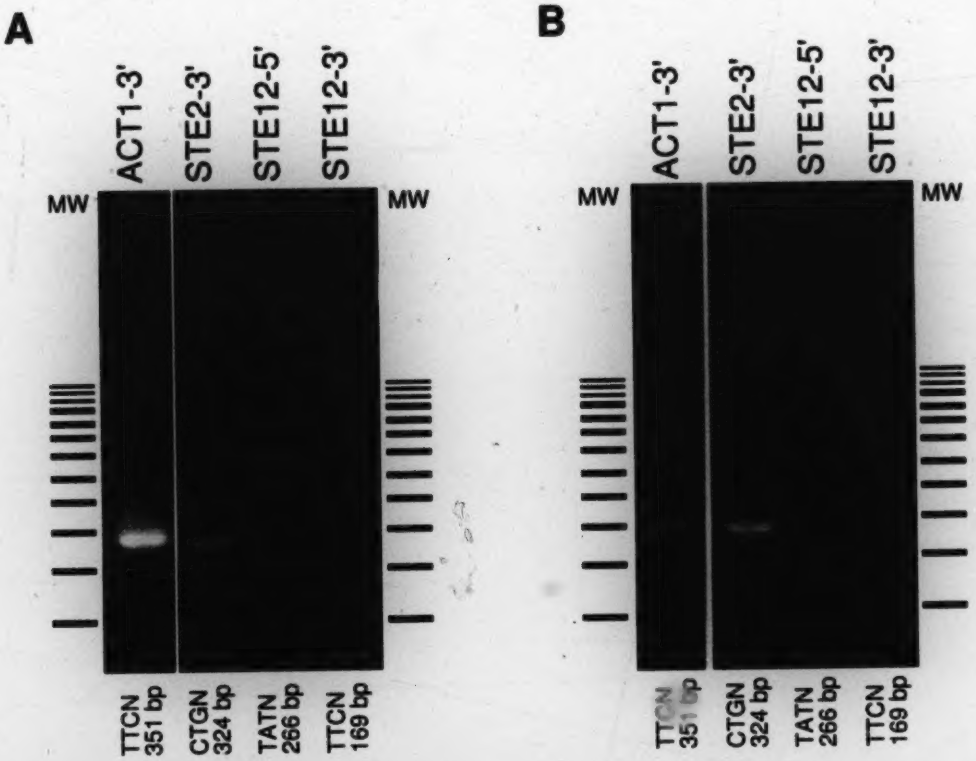


FIGURE 4.6: Amplification of indexed TF restriction fragments from *FokI*-digested yeast cDNA populations.
A) *FokI*-digested TF restriction fragments were indexed and amplified directly from a *FokI*-digested cDNA population. Each single-indexer-mix ligation reaction contained 20 ng of *FokI*-digested cDNA synthesized from a total cellular mRNA population isolated from cultures of *S. cerevisiae* strain W303 grown on glucose medium. Five hundred femtomoles (per indexer) of the appropriate BamCC NoP indexer mix was used to target and amplify the TF restriction fragments ACT1-3', STE2-3', STE12-5' or STE12-3'. These ds-cDNA transcripts were chosen for their ability to be specifically targeted by TF primers. Ligation was performed with 40 U of T4 DNA ligase in a 20-µl reaction volume for 14 hrs at 16°C. Two microlitres of each ligation were added to PCR reactions containing 20 pmol BamCC and 20 pmol of the appropriate TF directional primer. Amplification with *PfuTurbo*™ DNA polymerase was performed for 35 cycles using standard conditions. Amplified reactions were analyzed by agarose gel electrophoresis. B) *FokI*-digested TF fragments were ligated to biotinylated indexers, separated from unligated *FokI*-digested cDNA fragments by extraction with streptavidin-coated paramagnetic beads, and amplified using BamCC primer and appropriate directional TF primers. Ligations were performed as described above using 50 fmol (per indexer) of the appropriate biotinylated NoP indexer mix. Purification of indexer-ligated restriction fragments using Dynabeads was performed (as described in the text). Washed beads carrying the indexed templates were added to a 50-µl PCR reaction. Amplification with *Taq* DNA polymerase was performed for 35 cycles under standard PCR conditions for this enzyme. Amplified reactions were analyzed by agarose gel electrophoresis. Columns of black bars marked "MW" represent 123-bp ladder molecular weight markers.

sequence. The entire ligation reaction volume was added to a suspension of streptavidin-coated paramagnetic beads and incubated to permit binding of the biotinylated cDNA target fragments to the streptavidin. The beads, carrying cDNA restriction fragments ligated to indexers bearing the specific cohesive end sequence used in ligation, were added to a PCR reaction. Amplification of the specific target fragment set was facilitated by the reduced complexity of the cDNA restriction fragment population present in the PCR.

Ligations were performed in a similar manner to those described in **Section 4.3.4.1**, using 50 fmol (per indexer) of the appropriate NoP indexer mix. The reduced indexer concentration was employed to mitigate somewhat the forcing conditions employed in the previous experiment. The relative enrichment effect provided by paramagnetic bead purification of biotinylated restriction fragments was expected to compensate for the increased stringency of ligation.

Streptavidin-coated paramagnetic beads (Dynabeads M-280; Dynal AS, Oslo, Norway) were rinsed according to the supplier's instructions using a Dynal Magnetic Particle Concentrator (MPC). The rinsed beads were resuspended and 16 µl removed to a 200-µl PCR tube. The tube was placed for 3 minutes in a home-built magnetic particle concentrator rack using 1300 milliTesla NbFeB magnets (2 mm x 5 mm diam.) fused into the side of eight positions of a 96-well tube holder. The supernatant was pipetted off the beads, which were then suspended in 160 µl 2 M NaCl. Twenty microlitres of the washed Dynabead suspension were added to the 20-µl biotinylated-indexer ligations, resulting in a final hybridization concentration of 1 M NaCl. The biotinylated indexing ligation reactions were incubated with the Dynabead suspension at 37°C for 1 h with rotation. Following incubation, the Dynabeads were collected using the home-built MPC and the supernatant was removed. The beads were washed twice with 200 µl of 2 M NaCl prior to a final wash with ddH_2O and addition of the beads to an amplification reaction.

Amplification was performed in a similar manner to that outlined in the previous experiment. *Taq* DNA polymerase was employed as the amplifying enzyme, and standard conditions for PCR with this polymerase were used. Amplification products were analyzed by agarose gel electrophoresis (FIGURE 4.6B). Amplification of the four *Fok*I-digested TF terminal restriction fragments targeted in this investigation was observed. The ability to amplify target fragments from indexed templates attached to Dynabeads by biotin-streptavidin binding was demonstrated. The effect of reduced cDNA population complexity on amplification specificity was seen in the reduced number and level of amplification of spurious PCR products, as compared to the data presented in FIGURE 4.6A. However, the continued presence of spurious amplification products in several reactions indicated that improvements in Dynabead wash stringency were necessary to prevent non-specific DNA binding to the streptavidin-coated paramagnetic beads.

The target fragments amplified in this experiment carried an indexer at one end, and were primed by a transcript-specific primer from the other. The obvious next step in development of protocols for cDNA indexing of yeast was to perform a parallel set of amplification using the BamCC indexing primer and an anchored poly(T)$_{16}$-V primer, to target the poly(A) tails of 3'-end restriction fragments generated by *Fok*I digestion of the cDNA transcripts of interest. Amplifications of this type were assembled, containing target DNA from ligations performed in parallel with those from which the data of FIGURE 4.6B was generated. As *PfuTurbo*™ DNA polymerase exhibits 3'-5' proofreading activity, it was capable of removing the anchoring nucleotide of the poly(T)$_{16}$-V primers, and therefore could not be used for this application. *Taq* DNA polymerase was employed as the amplifying enzyme, due to its lack of 3'-5' proofreading activity. (*Taq* DNA polymerase had been used to generate the data presented in FIGURE 4.6B in an effort to provide continuity between that experiment and amplification of identical target templates using poly(T)$_{16}$-V primers in place of transcript-specific primers.) No amplification of target fragments, or of spurious PCR products, was observed using BamCC and poly(T)$_{16}$-V primers (data not shown).

Two possible explanations were investigated for the failure of poly(T)$_{16}$-V primers, in conjunction with BamCC, to amplify 3'-end cDNA fragments ligated to indexers. First, competition by poly(A)-tailed target fragments may have been dramatically increased due to the presence of large numbers of unindexed fragments also bearing poly(A) tails. As observed in FIGURE 4.6B, the presence of spurious amplification products in addition to amplified indexed target fragments indicated that nonspecific binding of nonbiotinylated DNA to Dynabeads was not eliminated by the Dynabead wash strategy suggested by the manufacturer. Moderate levels of nonspecific DNA binding would allow the import of sufficiently high numbers of unindexed fragments for poly(T) primer, preventing target fragment amplification. As a remedy, alternate Dynabead wash strategies were evaluated to identify wash conditions of sufficient stringency to entirely eliminate nonspecific binding of nonbiotinylated DNA to Dynabeads. A second possibility for amplification failure, the inability of poly(T)$_{16}$-V oligonucleotides to prime replication under the PCR conditions employed, was investigated subsequently.

4.3.5 Evaluation of Dynabead wash regimens

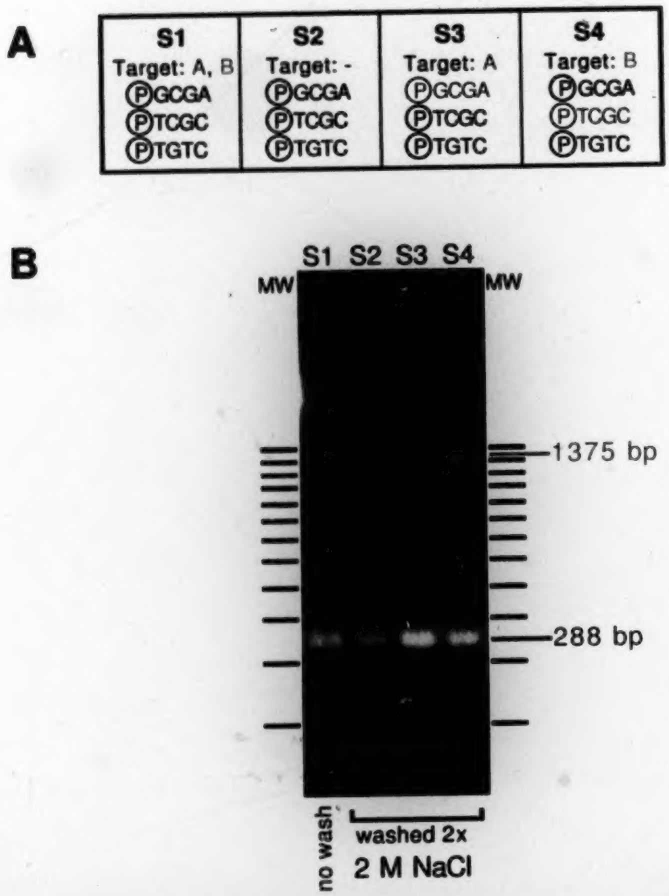
4.3.5.1 Evaluation of wash protocols to prevent non-specific binding of nonbiotinylated DNA fragments to streptavidin-coated paramagnetic beads

Particular combinations of biotinylated and nonbiotinylated indexers were ligated to pUC19 *Fok*I fragments to generate four sets of fragments with which to evaluate the stringency of several Dynabead wash strategies. A standard protocol previously described for the ligation of P-indexers to pUC19 was employed, with the modification that these ligations contained three P-indexers (at 50 fmol/targeting indexer/ligation) rather than the two P-indexers used in **Chapter II**. The combinations of biotinylated and unmodified indexers employed was designed to permit various combinations of fragments to be captured by streptavidin affinity from each ligation set, in the manner described below.

4.3.5.2 *Ligation and PCR of biotinylated pUC19 fragments for use in Dynabead wash regimen evaluations*

Four sets of ligation reactions were prepared for evaluating the success of different wash strategies in preventing non-specific binding of nonbiotinylated DNA fragments to streptavidin-coated paramagnetic beads: S1, S2, S3 and S4 (FIGURE 4.7A). Ligation S2 contained three phosphorylated nonbiotinylated indexers, targeting pUC19 fragments A and B. The absence of a biotinylated indexer was expected to prohibit the amplification of any fragment from Ligation S2 following Dynabead extraction, as effective washing of the Dynabeads would remove any templates bound to the paramagnetic beads in a non-specific manner. Ligation S3 contained the biotinylated indexer targeting the non-repeated end of the pUC19 A fragment, and two non-biotinylated phosphorylated indexers permitting simultaneous indexing of pUC19 B. Dynabead extraction of Ligation S3 was expected to permit the capture of biotinylated A fragment templates and remove the nonbiotinylated B fragment templates prior to amplification, if an effective wash strategy was used. Ligation S4 contained the biotinylated indexer targeting the non-repeated end of the pUC19 B fragment, and two non-biotinylated phosphorylated indexers permitting concomitant indexing of pUC19 A. Dynabead extraction of Ligation S4 was expected to permit the capture of biotinylated B fragment templates and remove the A templates prior to amplification, if an effective wash strategy was used. Ligation S1 was identical to Ligation S2, and was used only as a no-wash control for effective ligation and amplification.

To evaluate the usefulness of these fragment sets as a means of studying non-specific binding, a set of four 20-µl ligation were performed using indexer combination S1 to S4 and *Fok*I-digested pUC19 DNA. Following ligation, 20 µl of pre-rinsed Dynabead suspension was added to the reactions, which were then incubated for 1 hr at 37°C with agitation. A 4-µl aliquot of the unwashed Ligations S1 Dynabead suspension was added directly to a PCR reaction. For Ligations S2 to S4, the incubated Dynabead suspensions were each washed twice with 40 µl 2 M NaCl and



S1	S2	S3	S4
Target: A, B	Target: -	Target: A	Target: B
ⓅGCGA	ⓅGCGA	ⓅGCGA	ⓅGCGA
ⓅTCGC	ⓅTCGC	ⓅTCGC	ⓅTCGC
ⓅTGTC	ⓅTGTC	ⓅTGTC	ⓅTGTC

FIGURE 4.7: Amplification of biotinylated indexed pUC19 fragments for use in evaluating Dynabead wash regimens.
A) A series of four ligation reactions performed using the particular combinations of biotinylated and nonbiotinylated Bam series P-indexers outlined. Cohesive end sequences corresponding to biotinylated P-indexers present in indexer combinations S3 and S4 are noted in blue.
B) A set of four 20-µl ligations were performed using indexer combinations S1 to S4 and *Fok*I-digested pUC19 DNA according to standard protocols outlined in **Chapter II**. Twenty microlitres of washed Dynabead suspension was added to the complete reactions, and incubated with gentle agitation for 1 hr at 37°C. Two washes of the Dynabead suspension with 50 µl of 2 M NaCl were performed, corresponding to the supplier's instructions for removal of nonbiotinylated DNA. The washed beads were added directly to a 20-µl PCR reaction, amplified using *Taq* DNA polymerase, and the reaction products analyzed by agarose gel electrophoresis. Columns of black bars marked "MW" represent 123-bp ladder molecular weight markers.

rinsed with ddH_2O, according to the supplier's instructions. Twenty-microlitre aliquots of complete amplification reactions were added to the rinsed Dynabeads carrying DNA from ligations S2 to S4. Amplification of reactions S1 to S4 was performed as previously described for indexed pUC19 DNA fragments, and the PCR products were analyzed by agarose gel electrophoresis (FIGURE 4.7B).

pUC19 fragments A and B were amplified from the unwashed Dynabead suspension containing DNA from Ligation S1, providing a positive control for amplification of both indexed fragments. The amplification of both nonbiotinylated A and B fragment templates from washed Dynabeads incubated with Ligation S2 demonstrated that nonbiotinylated DNA was capable of nonspecific binding to streptavidin-coated beads, and that the supplier's recommended wash strategy was not sufficiently stringent to remove nonspecifically-bound DNA templates from the beads. This conclusion was supported by the results of amplification from washed Dynabeads incubated with Ligations S3 and S4. Although biotinylated fragments were captured by the streptavidin-coated beads (Fragment A in S3 and Fragment B in S4), nonbiotinylated fragments (Fragment B in S3 and Fragment A in S4) were bound to the beads in sufficient quantity as to be efficiently and competitively amplified with specifically-captured biotinylated target fragments. This experiment demonstrated the utility of the S1 to S4 fragment sets in determining Dynabead wash regimens with sufficient stringency to eliminate nonspecific binding of nonbiotinylated DNA fragments to streptavidin-coated paramagnetic beads.

4.3.5.3 Identification of wash regimens eliminating nonspecific DNA binding to streptavidin-coated paramagnetic beads

Nonspecific binding of nonbiotinylated DNA fragments to Dynabeads had been previously reported by Lund *et al.* [245], and alternate wash strategies to those recommended by the supplier had been suggested to reduce or eliminate this nonspecific binding [245, 246]. An evaluation of several wash regimens suitable for application to 3'-end cDNA indexing was performed using the S2, S3 and S4 fragment sets from the previous experiment. Two-hundred-microlitre ligation reactions were prepared for each of the S2, S3 and S4 indexer combinations. The contents of the ligation reactions were as described for the previous experiment, using identical reagent concentrations and scaling up the volumes and amounts used of each reagent ten-fold from those used in typical 20-µl ligation reactions. A single standard 20-µl reaction was prepared for Ligation S1 (data not shown). Ligation was performed at 37°C for 1 h with T4 DNA ligase, and the enzyme was denatured for 15 min at 65°C. Five hundred microlitres of Dynabead stock suspension was rinsed as suggested by the manufacturer and resuspended in 1000 µl 2 M NaCl. From each large-scale ligation, 200 µl was removed and added to an equal volume of prerinsed Dynabead suspension in a 1.5-ml microcentrifuge tube. The tubes were incubated for 2 h at 37°C with gentle agitation.

Following incubation, the contents of each Dynabead extraction (S2 to S4) were aliquoted into 7 separate 200-µl microcentrifuge tubes in preparation for the various wash regimens. For each Ligation S2 to S4, the washing strategies were as follows:

Tube:	**Wash Regimen**
i	1x with 500 µl ddH$_2$O
ii	2x with 500 µl ddH$_2$O
iii	1x with 500 µl 2 M NaCl
iv	2x with 500 µl 2 M NaCl
v	3x with 500 µl 2 M NaCl
vi	2x with 500 µl 5xSSC 2x with 500 µl 2xSSC 1x with 500 µl 2xSSC/50% formamide
vii	1x with 100 µl 5xSSC 1x with 100 µl 2xSSC/50% formamide

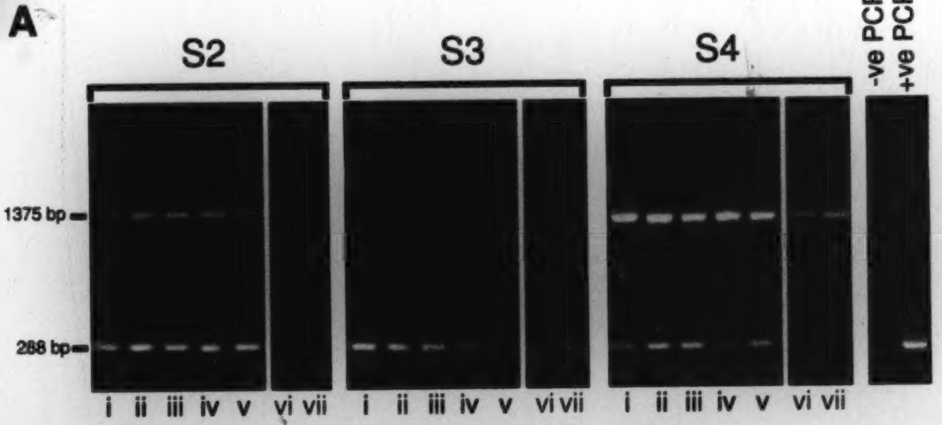
An additional rinse step with 100 µl ddH$_2$O was performed for Dynabeads in tubes iii to vii to remove residual salt prior to PCR.

A master indexing PCR reaction mix was prepared for each tube according to the standard protocol described for amplification of indexed pUC19 fragments by *Taq* DNA polymerase, in the absence of DNA template. Eighteen microlitres of this PCR reaction mix was aliquoted into each tube of unsuspended washed Dynabeads. Amplification was performed using a standard indexing PCR protocol. Analysis of PCR reactions was performed by agarose gel electrophoresis.

The data presented in FIGURE 4.8A indicated that neither higher wash volumes nor an increased number of wash steps significantly reduced the amount of amplified product generated from nonbiotinylated DNA fragments bound non-specifically to streptavidin-coated beads (washes i-v). A particularly stringent wash regimen [245] including several wash steps with high volumes of concentrated salt solutions in combination with formamide eliminated amplification of nonbiotinylated DNA fragments from Dynabeads incubated with both biotinylated and nonbiotinylated indexed targets (wash vi). A modified version of this regimen, featuring reduced numbers of wash steps and reduced wash volumes (wash vii), was also effective in eliminating nonspecific DNA binding. Additionally, this modified wash regimen was less labour-intensive and provided increased sample recovery, resulting in more efficient amplification from larger amounts of biotinylated indexed template. This wash regimen (100 µl 5xSSC; 100 µl 2xSSC/50% formamide; 100 µl ddH$_2$O) was selected as the most suitable Dynabead wash strategy for 3'-end cDNA indexing.

4.3.6 Attempted amplification of biotinylated indexed cDNA fragments from *Fok*I-digested cDNA populations with BamCC and anchored poly(T) primers following Dynabead extraction

The identification of a Dynabead wash strategy eliminating nonspecific DNA binding provided resolution to one of the two possibilities considered for the failure of indexed 3'-end cDNA fragments to amplify (discussed in **Section 4.3.4.2**). To observe



Tube:	Wash Regimen:
i	1x with 500 µl ddH$_2$O
ii	2x with 500 µl ddH$_2$O
iii	1x with 500 µl 2 M NaCl
iv	2x with 500 µl 2 M NaCl
v	3x with 500 µl 2 M NaCl
vi	2x with 500 µl 5xSSC 2x with 500 µl 2xSSC 1x with 500 µl 2xSSC/50% formamide
vii	1x with 100 µl 5xSSC 1x with 100 µl 2xSSC/50% formamide

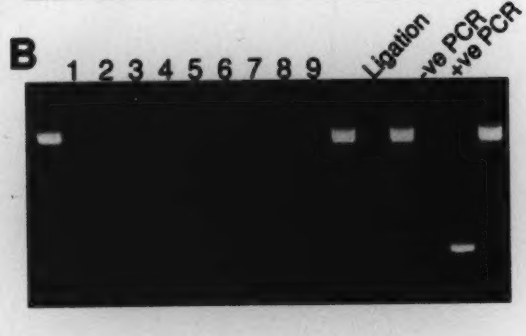


FIGURE 4.8: Evaluation of wash regimens for reduction of nonspecific DNA binding to streptavidin-coated paramagnetic beads.

A) Evaluation of several wash regimens was performed using the S2, S3 and S4 fragment sets. Ligations [200 µl total volume] were prepared for each of the S2, S3 and S4 indexer combinations, using ligation conditions and reagent concentrations identical to those used in FIGURE 4.7. Each completed ligation was added to an equal volume of prerinsed Dynabead suspension and incubated for 2 h at 37°C. Each extraction was aliquoted [20 µl/aliquot] into seven 200-µl microcentrifuge tubes in preparation for the wash regimens, which were then performed. An additional rinse step with 100 µl ddH$_2$O was performed for Dynabeads in tubes iii to vii. PCR reaction mix was added [18 µl/tube] to each tube of unsuspended washed Dynabeads. PCR and gel electrophoresis were performed as previously described.

B) A series of 3'-end cDNA indexing reactions were prepared from several *Fok*I-digested cDNA populations to evaluate whether a stringent Dynabead wash regimen would permit the successful amplification of biotinylated 3'-end cDNA fragments using BamCC and anchored poly(T) primers. Ligations 1 to 3 employed a *Fok*I-digested cDNA population derived from *S. cerevisiae* strain W303 grown in glucose medium in the presence of pheromone as target DNA. Ligations 4 to 6 targeted a *Fok*I-digested cDNA population derived from strain W303 grown in the absence of pheromone. Ligations 7 to 9 targeted a *Fok*I-digested cDNA population derived from strain yAO6 grown on glucose medium in the absence of pheromone. Ligations 1, 4 and 7 contained the NoP indexer mix OH-TTCNxBamCC, targeting the ACT1 and STE12 3'-end cDNA transcripts; ligations 2, 5 and 8 used OH-CTCNxBamCC to target the FUS1 3'-end cDNA transcript; and ligations 3, 6 and 9 employed OH-CTGNxBamCC to target the STE2 3'-end cDNA transcript. The ligation and wash-efficacy control involved the ligation of pUC19 by the indexer combination S3. Other details of this experiment are outlined in the text.

whether use of a stringent Dynabead wash regimen, eliminating nonspecific binding of competitive numbers of nonbiotinylated poly(A)-tailed transcripts, was sufficient to permit the successful amplification of biotinylated 3'-end cDNA fragments using BamCC and anchored poly(T) primers, a series of 3'-end cDNA indexing reactions were prepared from several *Fok*I-digested cDNA populations.

Ligation of *Fok*I-digested yeast ds-cDNA stocks $W303_{(glucose)}$, $W303_{(pheromone)}$, and $yAO6_{(glucose)}$ with biotinylated BamCC indexer mixes was performed in a manner similar to that employed in the indexing of Test Fragment cDNAs. Twenty nanograms of the appropriate *Fok*I-digested yeast ds-cDNA stock were added to a single-indexer ligation reaction containing 500 fmol of biotinylated nonphosphorylated BamCC indexer mix specific for the 3'-end *Fok*I fragment of either *ACT1*, *FUS1*, *STE2* or *STE12* transcripts.

A ligation and wash-efficacy control was included in this experiment that contained *Fok*I-digested pUC19 DNA, the biotinylated phosphorylated indexer specific for the non-repeated end of pUC19 A fragment, the non-biotinylated phosphorylated indexer specific for the non-repeated end of pUC19 B fragment, and the non-biotinylated nonphosphorylated indexer specific for the cohesive-end sequence common to both A and B. Amplification of this control was expected to generate only fragment A if the wash regimen used was successful in removing non-specifically-bound DNA fragments from the Dynabeads. If the wash did not prevent non-specific binding, both A and B would be amplified. If ligation failed, or if the template DNA was lost during the wash regimen, no product would be amplified.

Twenty microlitres of a standard washed Dynabead suspension were added to the completed 20-µl ligation reactions, and incubation was performed at 37°C for 2 hrs with tumbling rotation. The stringent washing regimen from the previous experiment was used to remove non-specifically-bound cDNA templates from the paramagnetic beads: the beads were washed once with 100 µl of 5x SSC, once with 100 µl of 2x SSC/ 50% formamide, and rinsed with 100 µl ddH_2O. An indexing PCR reaction mix was prepared for each tube in a manner similar to that described for the previous experiment. Eighteen microlitres of PCR reaction mix were aliquoted into each tube of unsuspended washed cDNA-bearing Dynabeads. Each PCR reaction, employing bead-bound indexed cDNA fragments as template, contained 20 pmol of BamCC indexing primer and 60 pmol of Poly(T)$_{35}$-V anchored primer mix (20 pmol/µl/oligo). The ligation/wash control contained 40 pmol of Bam indexing primer. A positive PCR control (a previously-amplified stock of pUC19 A fragment, diluted 1:10,000 and amplified using 40 pmol Bam primer) and a negative PCR control (20 pmol BamCC indexing primer and 60 pmol Poly(T)$_{35}$-V anchored primer mix in a PCR reaction devoid of target DNA) were also performed. Standard PCR conditions for the amplification of fragments by *Taq* DNA polymerase were employed, with 40 cycles of amplification. Amplified reactions were analyzed by agarose gel electrophoresis (FIGURE 4.8B).

The successful amplification of pUC19 fragment A alone as the ligation control in this experiment demonstrated that ligation of biotinylated indexers to target fragments had occurred. The specific capture of biotinylated indexed fragments by streptavidin-coated paramagnetic beads was accomplished. The removal of nonspecifically-bound nonbiotinylated DNA fragments from Dynabeads using a stringent washing regimen had also been successful, as evidenced by the absence of concomitant amplification of the nonbiotinylated pUC19 Fragment B from the ligation mixture following Dynabead capture and washing. No PCR contamination was present to outcompete legitimate amplification of indexed cDNA targets, as shown by the absence of PCR products in the negative PCR control. Additional confirmation that the PCR reactions were correctly assembled and had the potential to amplify DNA fragments was provided by the positive PCR control. The amplification of indexed *Fok*I-digested TF terminal restriction fragments had been previously demonstrated (FIGURE 4.6B), providing confirmation that these reaction conditions permitted indexers to be ligated to cDNA target fragments generating amplifiable target fragments. The inability of anchored poly(T) primers to efficiently initiate strand elongation in PCR of indexed 3'-end cDNA transcripts needed to be addressed.

4.3.7 Design, construction and use of FakePolyT model fragments in optimization of poly(T) priming strategies

4.3.7.1 Designing the BigPolyT Indexer for construction of FakePolyT fragments

In order to investigate strategies enabling the amplification of indexed 3'-end cDNA fragments by BamCC and anchored poly(T) primers and to facilitate troubleshooting and optimization of cDNA indexing amplification protocols, it was necessary to generate a set of artificial templates (*FakePolyT*s) mimicking features of indexed 3'-end cDNA fragments. A large double-stranded oligonucleotide that would facilitate the formation and amplification of a FakePolyT molecule was designed and synthesized (FIGURE 4.9A). This double-stranded oligo, referred to as *BigPolyT Indexer*, was composed of the two complementary oligonucleotides *BamPolyTPrimer* and *PolyA-TGTCindexer*. When annealed, these two oligos could be used as a large double-stranded indexer, capable of targeting and ligating to both ends of the pUC19 AB double fragment (or one end of either the pUC19 A or B fragments), being primed and amplified by Bam indexing primer in a standard indexing amplification reaction, and subsequently acting as a template for poly(T) priming optimization.

4.3.7.2 Approach 1: Ligation of BigPolyT Indexer to pUC19 A & B fragments

Artificial indexed 3'-end cDNA templates were constructed by ligation of BigPolyT Indexer directly to Fragment A and Fragment B of pUC19 (FIGURE 4.9B). Four ligation reactions were prepared, two to ligate BigPolyT Indexer to the *Fok*I-digested pUC19 fragments and two to act as ligation and amplification controls for their construction. Tube 1 contained 1 ng of *Fok*I-digested pUC19 DNA, 50 fmol of BigPolyT Indexer, and 50 fmol/oligo of OH-GCGNxBamCC indexer mix in a standard ligation reaction. This ligation was designed to construct a 325-bp fragment (A^+) exhibiting a BamCC primer-binding site at one end and at the other a 35-nt stretch of thymidine bases immediately downstream of a Bam primer-binding site. Tube 2 was similar to Tube 1, with the substitution of indexer mix to 50 fmol/oligo OH-TCGNxBamCC. This ligation was designed to construct a 1412-bp fragment (B^+) with

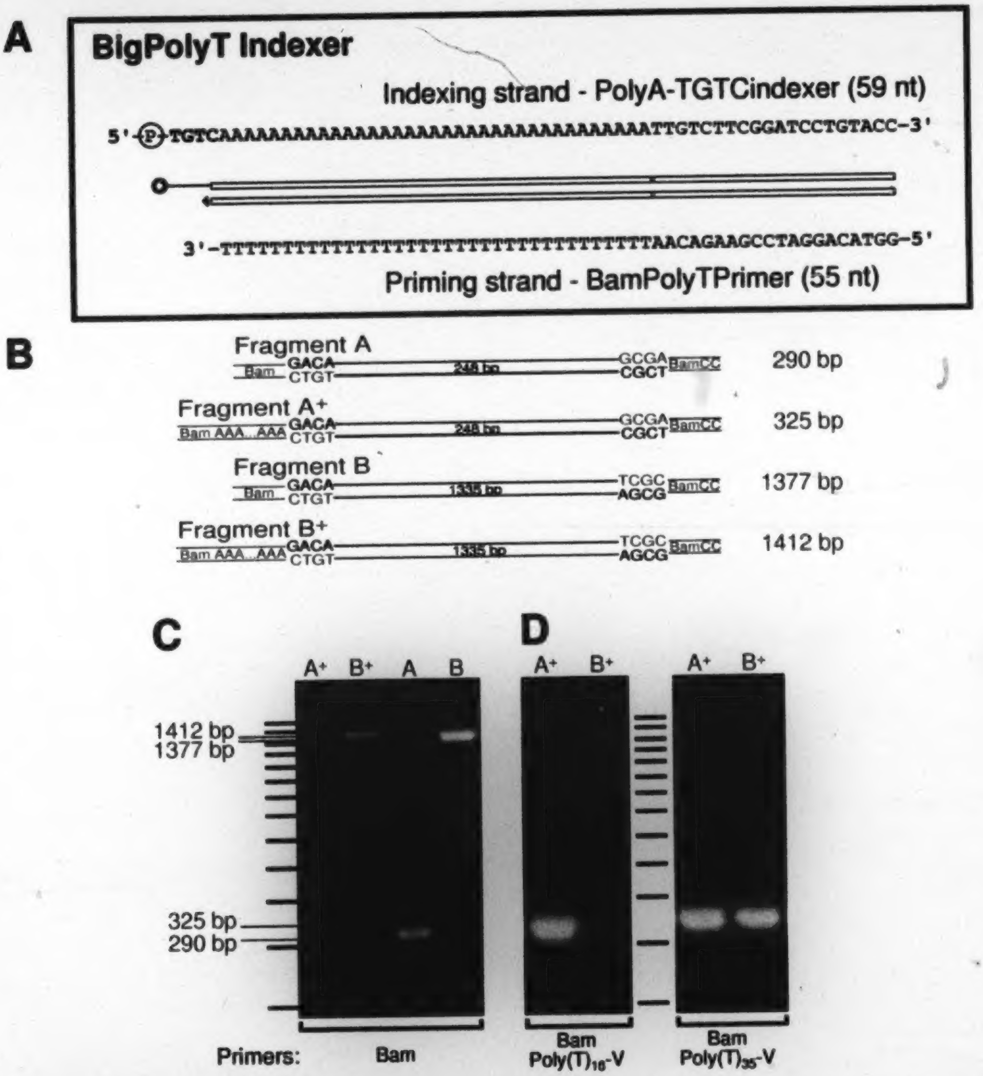

A
BigPolyT Indexer
Indexing strand - PolyA-TGTCindexer (59 nt)
5'-P-TGTCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAATTGTCTTCGGATCCTGTACC-3'
3'-TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTAACAGAAGCCTAGGACATGG-5'
Priming strand - BamPolyTPrimer (55 nt)
B
Fragment A
Bam GACA CTGT 248 bp GCGA CGCT BamCC 290 bp
Fragment A+
Bam AAA...AAA GACA CTGT 248 bp GCGA CGCT BamCC 325 bp
Fragment B
Bam GACA CTGT 1335 bp TCGC AGCG BamCC 1377 bp
Fragment B+
Bam AAA...AAA GACA CTGT 1335 bp TCGC AGCG BamCC 1412 bp
C
A+ B+ A B
1412 bp
1377 bp
325 bp
290 bp
Primers:
Bam
D
A+ B+
A+ B+
Bam Poly(T)16-V
Bam Poly(T)35-V

FIGURE 4.9: Design, construction and use of FakePolyT model fragments in optimization of poly(T) priming strategies.

A) Sequence of BigPolyT Indexer oligonucleotides.

B) Structure of A$^+$ and B$^+$ fragments compared to that of indexed pUC19 Fragments A and B.

C) Amplification of A$^+$ and B$^+$ fragments in preparation for gel purification. Indexed pUC19 Fragments A and B were also amplified for comparison. See text for further details.

D) Amplification of gel-purified A$^+$ and B$^+$ fragments using BamCC and Poly(T)$_{16}$-V primers was compared to the amplification of these fragments with BamCC and Poly(T)$_{35}$-V. Experimental details are outlined in the text.

features similar to those of Tube 1. Tubes 3 and 4 were similar to Tubes 1 and 2, respectively, with the substitution of 50 fmol of P-TGTCxBam for BigPolyT Indexer. These ligation controls indexed the pUC19 A and B fragments, generating indexing templates of 290 bp and 1377 bp, respectively. A standard indexing protocol was used for ligations. Amplification of the ligated samples with 40 pmol Bam primer was performed by *PfuTurbo*™ DNA polymerase in a standard PCR reaction. Analysis of amplified fragments was performed by agarose gel electrophoresis (FIGURE 4.9C). Slices of agarose containing the A^+ and B^+ fragments were cut from the gel, cleaned, resuspended and diluted to 1 pg/μl in ddH_2O.

4.3.7.3 Comparison of amplification efficiency of 3'-primer sets Poly(T)$_{35}$-V and Poly(T)$_{16}$-V

Amplification of the gel-purified A^+ and B^+ fragments using BamCC primer and Poly(T)$_{35}$-V was compared to amplification of these fragments with BamCC and Poly(T)$_{16}$-V (FIGURE 4.9D). Standard PCR reactions containing 20 pmol BamCC and 20 pmol/oligo (60 pmol total) anchored poly(T) primers were prepared with *Taq* DNA polymerase. Amplification proceeded for 40 cycles using cycling times and temperatures described for standard indexing PCRs. Evaluation of anchored poly(T) primer amplification efficiency was performed by agarose gel electrophoresis analysis of the amplified reactions.

Both Poly(T)$_{16}$-V and Poly(T)$_{35}$-V primers were able to amplify the A^+ FakePolyT fragment from gel-isolated A^+ template in combination with BamCC primer. However, amplification of the B^+ FakePolyT fragment was not accomplished efficiently using either anchored poly(T) primer. In the case of Poly(T)$_{16}$-V, weak amplification of the B^+ template was accompanied by a high level of nonspecific ("background") amplification producing a smear on the gel, in addition to concomitant weak amplification of the A^+ template. Amplification of gel-isolated B^+ template by Poly(T)$_{35}$-V and BamCC was completely outcompeted by the undesirable amplification of A^+ template. It was determined that preparations of gel-isolated B^+ template were contaminated by sufficiently high amounts of A^+ template to permit competitive amplification of A^+. As a result, this approach to the construction of FakePolyT fragments for 3'-end cDNA indexing protocols was discontinued. Because of the high level of background generated by amplification of simple targets with Poly(T)$_{16}$-V, and because of the significant dissimilarity in annealing temperature between the Poly(T)$_{16}$-V and BamCC oligonucleotides, no further development of cDNA indexing protocols were performed with using the anchored Poly(T)$_{16}$-V primers.

4.3.7.4 Design of GCRichPoly(T)$_{16}$-V primer

A poly(T) primer was designed that had a T_m well matched to that of the BamCC primer (unlike Poly(T)$_{16}$-V) but contained a significantly shorter poly(T) region than Poly(T)$_{35}$-V, reducing the potential negative effects on amplification efficiency produced as a result of secondary structure of a homopolymeric oligonucleotide primer. The anchored poly(T) primer variant GCRichPoly(T)$_{16}$-V (5'-GGGCACGC(T)$_{16}$V-3') featured an eight-base 5' GC-rich region, a 16-nt poly(T) stretch for priming first-strand cDNA synthesis from mRNA templates and for amplifying 3'-end cDNA fragments, and a single anchoring nucleotide (A, C or G). Initiation of first-strand cDNA synthesis with this oligo set was anticipated to provide increased priming specificity. Additionally, the new anchored poly(T) primer set was expected to be more efficient in amplifying indexed 3'-end cDNA fragments generated from mRNA populations using these oligos during cDNA synthesis.

4.3.7.5 Approach 2: Construction of a DNA target amplicon primed at both ends by poly(T) primers

Attempts to construct a set of FakePolyT fragments by direct ligation of BigPolyT Indexer and BamCC indexers to pUC19 *Fok*I restriction fragments A and B failed. This failure was due to the difficulty of completely purifying the B^+ template stocks from all traces of the A^+ template. Additionally, the presence of the Bam primer binding sequence on both ends of both templates in the initial ligation and amplification steps of their construction made it difficult to ensure that BamCC was not capable of illegitimately priming from these sequences at the low temperatures required for Poly(T)$_{16}$-V annealing during the annealing step of PCR cycling. A second approach to FakePolyT construction was designed to avoid these difficulties. This approach involved the ligation of BigPolyT Indexer to both ends of a suitable restriction fragment; amplification of the fragment using only anchored poly(T) primers with resultant loss of the Bam-primer binding sequences on the poly(T) amplicon; *Fok*I digestion of the poly(T) amplicon into two restriction fragments bearing indexable cohesive end sequences on one of their ends and a poly(A)/poly(T) tail at the other; ligation of BamCC NoP indexer mixes to these indexable ends; and preparative amplification of the completely constructed FakePolyT fragments for use as templates for cDNA indexing protocol development.

The *Fok*I-digested pUC19 AB double fragment (see **Section 2.3.4**) was selected as a suitable target for this approach to FakePolyT construction. The *Fok*I recognition sequence of the restriction site generating the GACA cohesive end of Fragment A is located upstream of its cutsite. The *Fok*I recognition sequence of the restriction site generating the GACA cohesive end of Fragment B is located downstream of the cutsite. In other words, the recognition sequences of the restriction sites generating the AB double fragment are external to the sequence of the double fragment itself. Ligation of BigPolyT Indexer (bearing the cohesive end sequence TGTC), or P-TGTCxBam indexer to the AB double fragment results in a fragment amplifiable using Bam primer alone. This indexable fragment contains a single *Fok*I recognition sequence and restriction site - the internal *FokI* site generating the A and B fragments from the AB double fragment. Digestion of the indexed AB double fragment does not remove the attached indexers, but cleaves the double fragment into indexable A and B fragments carrying indexer sequences of their distal ends. FakePolyT construction proceeded using the AB double fragment as a target (FIGURE 4.10).

BigPolyT Indexer was ligated to both GACA cohesive ends of the *Fok*I-digested pUC19 AB double fragment, which required an additional ligation event to rejoin the adjacent cohesive end sequences of Fragment A and Fragment B. (The rejoining of A and B fragments occurred in the same period of ligation as the attachment of BigPolyT Indexer.) Assembly of conventionally-indexed AB double

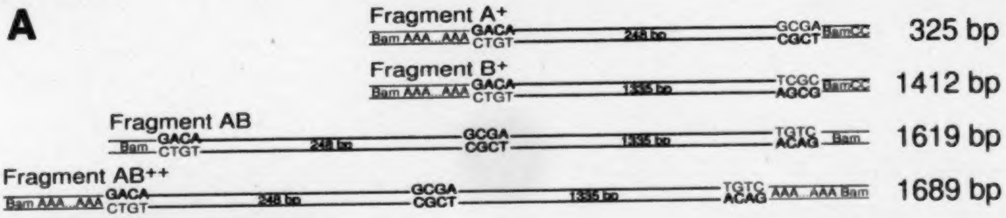
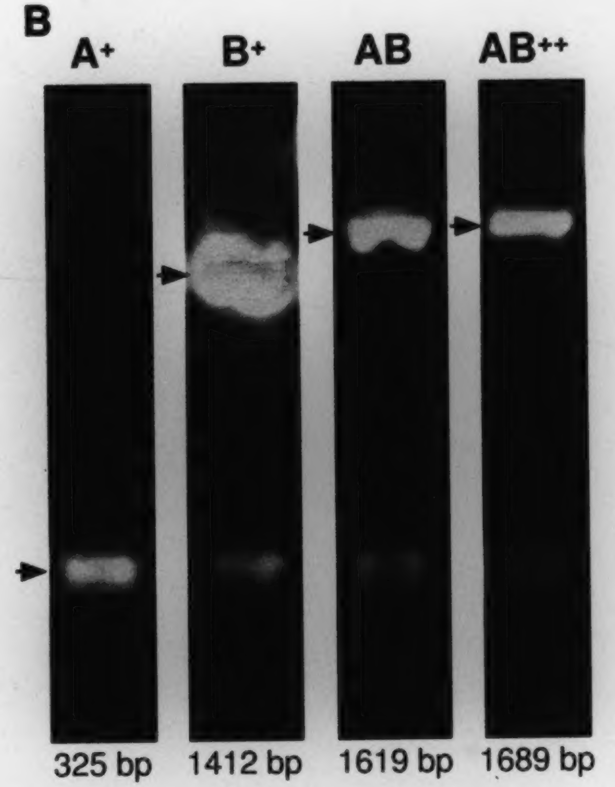

A
Fragment A+
Bam AAA...AAA
GACA
CTGT
248 bp
GCGA
CGCT
BamCC
325 bp
Fragment B+
Bam AAA...AAA
GACA
CTGT
1335 bp
TCGC
AGCG
BamCC
1412 bp
Fragment AB
Bam
GACA
CTGT
248 bp
GCGA
CGCT
1335 bp
TGTC
ACAG
Bam
1619 bp
Fragment AB++
Bam AAA...AAA
GACA
CTGT
248 bp
GCGA
CGCT
1335 bp
TGTC
ACAG
AAA...AAA Bam
1689 bp

FIGURE 4.10: Construction of FakePolyT amplicons primed at both ends by poly(T) primers.

A) The AB^{++} construct was assembled by ligation of BigPolyT Indexer to both terminal cohesive ends of the pUC19 AB double fragment, generating an indexed template 1689 bp in length. The AB double fragment was constructed in a parallel ligation using **P**-TGTCxBam (generating a 1619-bp indexed template), as a control for AB assembly and indexing. The A$^+$ and B$^+$ constructs are shown for comparison. All constructs assembled using BigPolyT Indexer are shown in pink, while the indexed AB construct generated using a conventional Bam P-indexer is illustrated in red.


B
A+
B+
AB
AB++
325 bp
1412 bp
1619 bp
1689 bp

B) Ligations contained 1 ng of *Fok*I-digested pUC19 DNA, 1 pmol of BigPolyT Indexer, standard ligase buffer components and 40 U T4 DNA ligase. Ligation was performed at 16°C overnight to ensure ligation of BigPolyT Indexer to each terminal cohesive end of the targeted AB double fragment. Amplification of the ligated samples in 50-µl reaction volumes with 40 pmol Bam primer was performed by *Taq* DNA polymerase for 35 cycles at standard cycling temperatures and times. Analysis of amplified AB and AB^{++} fragments was performed by agarose gel electrophoresis. Aliquots of the A$^+$ and B$^+$ fragments constructed and amplified in FIGURE 4.9D were electrophoresed in parallel with the AB and AB^{++} amplicons for comparison. Concomitant amplification of the A$^+$ construct was observed in all PCR reactions.

fragment was performed in parallel with the construct-assembly ligations, to provide a ligation and amplification control for the assembly process. In order to confirm assembly of the BigPolyT-indexed construct, and for comparison with the conventionally-indexed AB control, amplification of the ligated samples with 40 pmol Bam primer was performed. The Bam-primed amplification products were analyzed by agarose gel electrophoresis (FIGURE 4.10B). Successful amplification of the pUC19 AB double fragment with BigPolyT Indexer ligated to both ends (the AB^{++} construct) was observed. Amplification of the conventionally-indexed AB double fragment (referred to as the AB or *AB=Bam* construct) was also successful.

Amplification of AB^{++} construct with anchored poly(T) primers

Following confirmation of AB^{++} assembly by Bam-primed amplification, AB^{++} ligation products were amplified using either Poly(T)$_{35}$-V or GCRichPoly(T)$_{16}$-V as the sole primer (FIGURE 4.11A). (As a control, the indexed AB template was amplified in parallel using Bam primer.) The amplifications of AB^{++} with Poly(T)$_{35}$-V and GCRichPoly(T)$_{16}$-V generated two new constructs, AB^{++}=Poly(T)$_{35}$ and AB^{++}=GCRichPoly(T)$_{16}$, respectively. Both new amplicons lacked the terminal Bam primer binding sequences present on both ends of the AB^{++} construct. The AB^{++}=Poly(T)$_{35}$ amplicon carried terminal sequences consisting of 35mer tracts of poly(A)/poly(T) on each end. The AB^{++}=GCRichPoly(T)$_{16}$ amplicon carried terminal sequences consisting of an 8-bp GC-rich sequence external to a 16mer poly(A)/poly(T) tract on each end of the construct. In both cases, the terminal sequences of the amplicons mimicked the terminal sequences generated by the priming of first-strand cDNA synthesis from the poly(A) tails of mRNA transcripts with the appropriate anchored poly(T) primer.

Digestion of poly(T)-primed AB^{++} amplicons

Following their initial amplification, AB^{++}=Poly(T)$_{35}$ and AB^{++}=GCRichPoly(T)$_{16}$ were prepared for *Fok*I digestion. Slices of agarose containing one or other poly(T)-primed AB^{++} amplicon were cut from the gel, cleaned

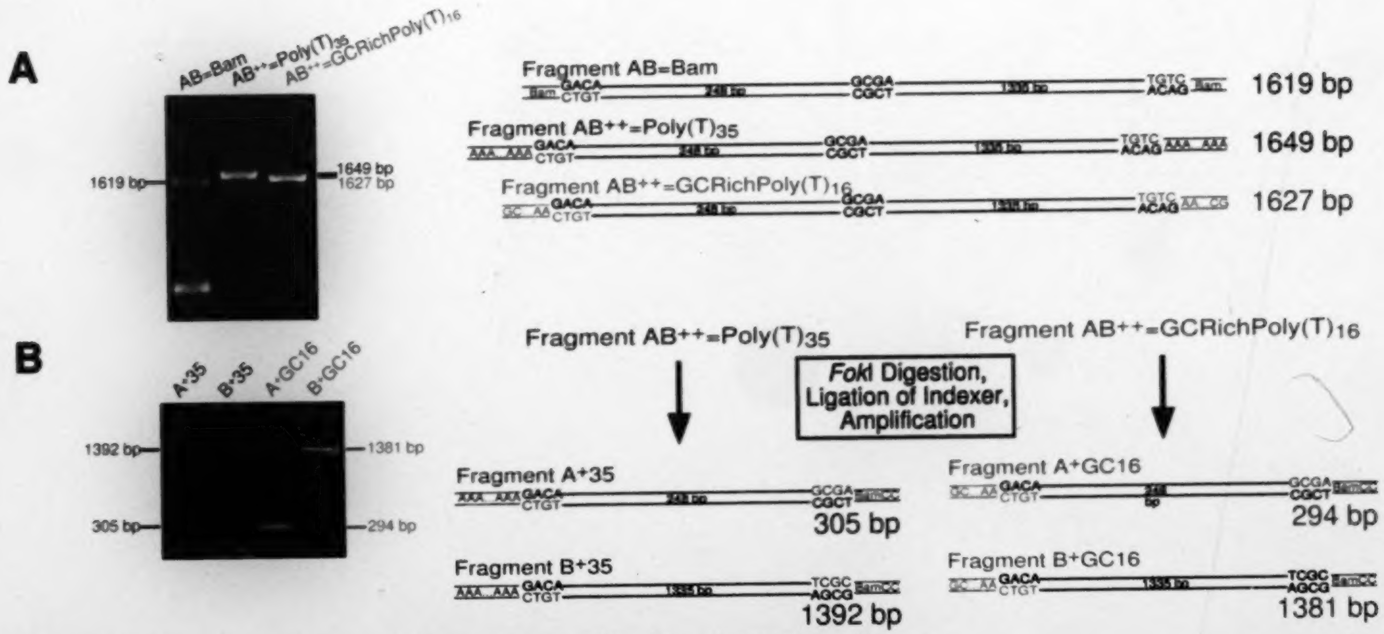

A
AB=Bam
AB++=Poly(T)35
AB++=GCRichPoly(T)16
1619 bp
1649 bp
1627 bp
Fragment AB=Bam
Bam
GACA
CTGT
GCGA
CGCT
TGTC
ACAG
1619 bp
Fragment AB++=Poly(T)35
AAA...AAA
1649 bp
Fragment AB++=GCRichPoly(T)16
GC...AA
AA...CG
1627 bp
B
A+35
B+35
A+GC16
B+GC16
1392 bp
305 bp
1381 bp
294 bp
Fragment AB++=Poly(T)35
Fragment AB++=GCRichPoly(T)16
FokI Digestion, Ligation of Indexer, Amplification
Fragment A+35
BamCC
305 bp
Fragment A+GC16
294 bp
Fragment B+35
TCGC
AGCG
1335 bp
1392 bp
Fragment B+GC16
1381 bp

FIGURE 4.11: Indexing of *FokI*-digested poly(T)-primed AB^{++} amplicons.
A) The AB^{++} construct was amplified using the anchored poly(T) primer mixes: Poly(T)$_{35}$-V and GCRichPoly(T)$_{16}$-V. Amplification of the construct was performed by *Taq* DNA polymerase using 40 pmol of a particular anchored poly(T) primer in a standard 50-µl PCR reaction for 35 cycles. The indexed AB template was amplified in parallel using Bam primer. PCR products were analyzed by agarose gel electrophoresis.

B) AB^{++} fragments amplified with Poly(T)$_{35}$-V and with GCRichPoly(T)$_{16}$-V were sliced from agarose gel, purified and resuspended in 100 µl ddH$_2$O, quantitated by UV spectrophotometry, and digested with *FokI* endonuclease. Fragments generated by *FokI* digestion of AB^{++}=Poly(T)$_{35}$ were referred to as A^{+}35 and B^{+}35, while fragments generated by *FokI* digestion of AB^{++}=GCRichPoly(T)$_{16}$ were referred to as A^{+}GC16 and B^{+}GC16. Digestion was confirmed by agarose gel electrophoresisis. *FokI*-digested AB^{++}=Poly(T)$_{35}$ and *FokI*-digested AB^{++}=GCRichPoly(T)$_{16}$ stocks were diluted to 1 ng/µl. Each of the four *FokI*-digested poly(T)-primed AB^{++}-derived constructs were ligated to appropriate BamCC NoP indexers. To generate indexed A^{+}35, 1 ng of *FokI*-digested AB^{++}=Poly(T)$_{35}$ was ligated to 50 fmol **OH**-GCGNxBamCC under standard indexing ligation conditions. B^{+}35 was generated in a similar manner, using **OH**-TCGNxBamCC as the indexer. A^{+}GC16 and B^{+}GC16 were generated in similar fashion, employing *FokI*-digested AB^{++}=GCRichPoly(T)$_{16}$ as the DNA source in both ligations. In subsequent iterations of this process, biotinylated BamCC NoP indexer mixes were used as appropriate for the task required. Amplification of FakePolyT fragments was performed in standard indexing PCRs using BamCC and the appropriate anchored poly(T) oligo mix as primers, and analyzed by agarose gel electrophoresis.

of impurities using the Concert™ PCR purification system, resuspended in 100 µl ddH$_2$O, and quantitated by UV spectrophotometry. In each of the two *Fok*I digests, 1 µg of the appropriate DNA construct was cleaved with 1 U *Fok*I endonuclease in a 30-µl digest. The digests were incubated for 1 hr at 25°C, the enzyme heat-denatured at 65°C, and the results of the digests analyzed by agarose gel electrophoresis (data not shown). The two fragments generated by *Fok*I digestion of AB^{++}=Poly(T)$_{35}$ were referred to as A$^+$35 and B$^+$35. The two fragments generated by *Fok*I digestion of AB^{++}=GCRichPoly(T)$_{16}$ were referred to as A$^+$GC16 and B$^+$GC16. Stocks of *Fok*I-digested AB^{++}=Poly(T)$_{35}$ and *Fok*I-digested AB^{++}=GCRichPoly(T)$_{16}$ were diluted to 1 ng/µl.

Ligation of BamCC NoP indexers to Fok*I-digested poly(T)-primed AB^{++} constructs for completion of FakePolyT assembly*

For the final step of constructing a set of FakePolyT templates capable of modeling the behaviour of indexed 3'-end cDNA fragments, each of the four fragments generated by *Fok*I digestion of poly(T)-primed AB-derived constructs were ligated to nonphosphorylated BamCC indexers. To generate indexed A$^+$35, one nanogram of *Fok*I-digested AB^{++}=Poly(T)$_{35}$ was ligated to 50 fmol OH-GCGNxBamCC under standard indexing ligation conditions. B$^+$35 was generated in a similar manner, substituting OH-TCGNxBamCC as the indexer. A$^+$GC16 and B$^+$GC16 were generated in similar fashion, substituting *Fok*I-digested AB^{++}=GCRichPoly(T)$_{16}$ as the DNA source in both ligations. In subsequent iterations of this process, biotinylated nonphosphorylated BamCC indexer mixes were used as appropriate for the task required.

Amplification of FakePolyT fragments was performed using standard indexing PCR techniques using BamCC and the appropriate anchored poly(T) oligo mix as primers. A control PCR reaction, employing a FakePolyT fragment as the DNA template and BamCC as the only primer, was used to ensure that the fragments were constructed in such a way that both BamCC and an anchored poly(T) oligo were required to prime PCR (data not shown). Successful amplification of all four FakePolyT fragments was confirmed by agarose gel electrophoresis (FIGURE 4.11B).

4.3.8 Use of FakePolyT fragments in development of cDNA indexing protocols

4.3.8.1 Comparison of efficiency of Poly(T)$_{35}$-V and GCRichPoly(T)$_{16}$-V primers for indexed 3'-end cDNA fragment amplification

The four FakePolyT fragments A$^+$35, B$^+$35, A$^+$GC16 and B$^+$GC16 were used to evaluate the comparative amplification efficiencies of the Poly(T)$_{35}$-V and GCRichPoly(T)$_{16}$-V primers. This experiment was designed to provide insight into both the optimal conformation of cDNA 3'-terminal sequences generated during first-strand cDNA synthesis, and for selection of anchored poly(T) primer for PCR of indexed 3'-end cDNA fragments. Ligations of the four FakePolyT fragments were added without Dynabead extraction to PCR reactions containing Bam primer and the relevant anchored poly(T) primer. Amplification was performed using standard indexing protocols for PCR of Dynabead-captured cDNA templates and the results analyzed on agarose gel (FIGURE 4.12).

Use of Poly(T)$_{35}$-V primer in combination with Bam primer failed to amplify A$^+$35 or A$^+$GC16 templates. The large B$^+$35 and B$^+$GC16 templates were inefficiently amplified to barely detectable levels on EtBr-stained agarose gel using this primer set. Despite the matched T_m of the Poly(T)$_{35}$-V primer with that of the Bam sequence, amplification of target was poor or nonexistent under cDNA indexing PCR conditions, regardless of the 3'-terminal sequences of the target. These results, which correlated with the failure of indexed 3'-end cDNA fragments to amplify in FIGURE 4.8B, indicated that Poly(T)$_{35}$-V was ineffective as a primer for indexed 3'-end cDNA fragment amplification. Additionally, these data suggested that ds-cDNA populations primed in first-strand cDNA synthesis with this primer would not provide efficient targets for 3'-end cDNA indexing.

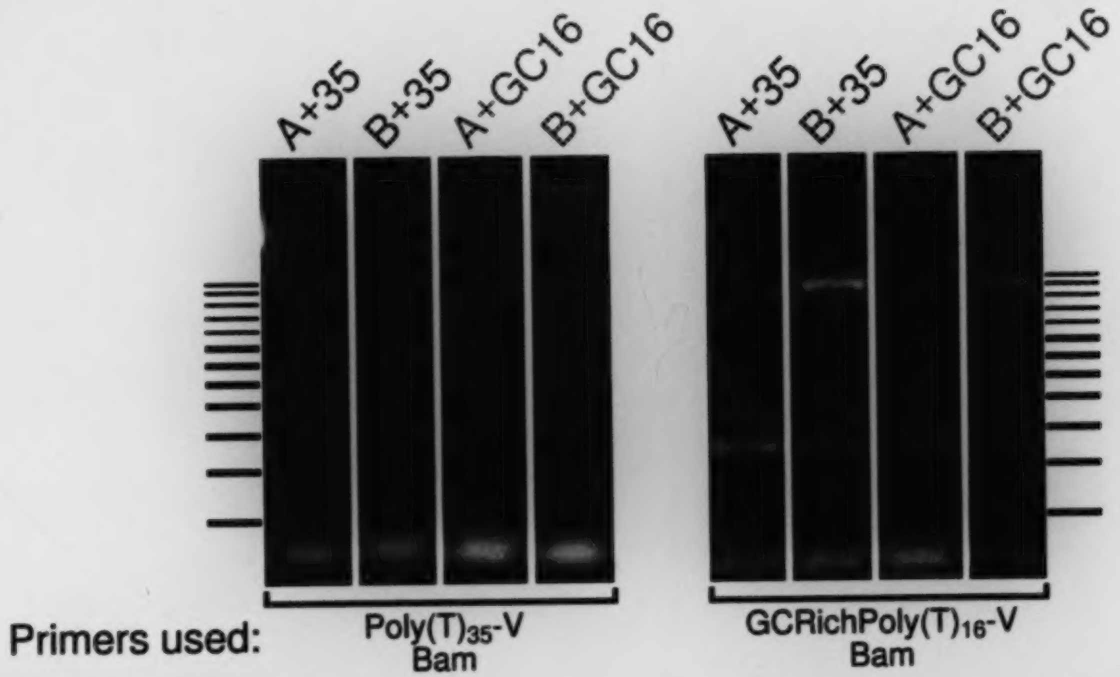

A+35
B+35
A+GC16
B+GC16
A+35
B+35
A+GC16
B+GC16
Primers used:
Poly(T)35-V
Bam
GCRichPoly(T)16-V
Bam

FIGURE 4.12: Comparison of efficiency of Poly(T)$_{35}$-V and GCRichPoly(T)$_{16}$-V primers for amplification of indexed FakePolyT fragments.
Ligations of $A^{+}35$, $B^{+}35$, $A^{+}GC16$ and $B^{+}GC16$ were performed in an identical manner to ligations described in FIGURE 4.11B. Biotinylated indexers were employed for continuity with subsequent experiments. No Dynabead extraction was performed. Two microlitres of the appropriate ligation was added to a 20-µl PCR reaction containing 40 pmol Bam primer and 40 pmol of either Poly(T)$_{35}$-V or GCRichPoly(T)$_{16}$-V. Thirty-five cycles of the PTC program "NEW55-62" were performed, and the PCR products were analyzed on agarose gel.

In contrast, the GCRichPoly(T)$_{16}$-V primer, in combination with Bam primer, provided efficient amplification of both A- and B-based templates with either type of 3'-terminal sequence structure. The amplification efficiency provided by this primer was sufficient to permit amplification of rare misligated fragment templates in this set of PCRs of ligated material unpurified by streptavidin capture on paramagnetic beads. GCRichPoly(T)$_{16}$-V was shown to be an efficient anchored poly(T) primer for the amplification of indexed 3'-end cDNA fragments.

4.3.8.2 Efficiency of 3'-end cDNA fragment amplification from Dynabead-bound biotinylated templates

The efficiency of amplification from streptavidin-purified indexed templates carrying 3'-terminal sequences complementary to the anchored poly(T) primer used in PCR was evaluated for Poly(T)$_{35}$-V and GCRichPoly(T)$_{16}$-V. Ligations of FakePolyT fragments were performed as in the previous experiment. Completed ligations were incubated with Dynabeads and indexed fragments isolated by streptavidin capture. Dynabead-bound indexed fragments were amplified in PCR reactions employing Bam primer and the appropriate anchored poly(T) primer. Ligations of $A^{+}35$ and $B^{+}35$ were amplified using Bam and Poly(T)$_{35}$-V, while A^{+}GC16 and B^{+}GC16 were amplified using Bam GCRichPoly(T)$_{16}$-V. Amplification products were analyzed on agarose gel (FIGURE 4.13A).

Poly(T)$_{35}$-V failed to amplify the A+35 template bearing a 3'-terminal 35mer poly(A)/poly(T) tract. Efficient amplification of both A^{+}GC16 and B^{+}GC16 following Dynabead purification was observed for the GCRichPoly(T)$_{16}$-V and Bam primer combination. This application of FakePolyT fragments clearly indicated the utility of GCRichPoly(T)$_{16}$-V as a primer for first-strand cDNA synthesis and for PCR in 3'-end cDNA indexing protocols.

4.3.8.3 Evaluation of PCR cycling parameters for 3'-end cDNA indexing

The amplification efficiencies provided by several PCR programs with slight variations in PTC cycling parameters were evaluated using the B^{+}GC16 and A^{+}GC16

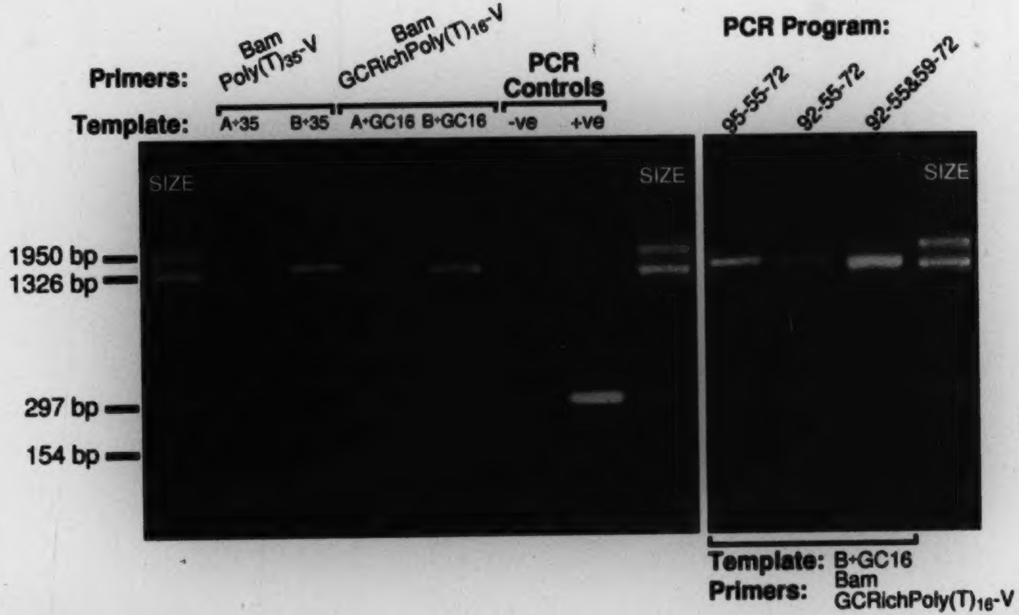

Primers:
Bam Poly(T)35-V
Bam GCRichPoly(T)16-V
PCR Controls
Template:
A+35
B+35
A+GC16
B+GC16
-ve
+ve
SIZE
SIZE
SIZE
1950 bp
1326 bp
297 bp
154 bp
PCR Program:
95-55-72
92-55-72
92-55&59-72
Template: B+GC16
Primers: Bam GCRichPoly(T)16-V

FIGURE 4.13: Amplification of biotinylated indexed artificial 3'-end cDNA fragments following streptavidin capture.
A) Ligations of FakePolyT templates with biotinylated indexers were performed as in FIGURE 4.12. Twenty microlitres of rinsed Dynabead suspension was added to each ligation and incubated for 1 hr at 37°C with agitation. Dynabeads were washed using the standard stringent wash regimen and added to 20-µl PCR reactions. Each PCR reaction contained 40 pmol Bam primer and 40 pmol of the appropriate anchored poly(T) primer. A negative PCR control containing no target DNA was performed, using 40 pmol Bam primer and 40 pmol GCRichPoly$(T)_{16}$-V. A positive PCR control containing the same primers in addition to 2 µl of a 1:10 000 dilution of a previous A+GC16 amplification reaction was employed. PCR proceeded for 35 cycles of PTC program "95-55-72", using the cycling parameters 30 s at 95°C; 30 s at 55°C; 60 s at 72°C. Analysis of target amplifications and control reactions was performed by agarose gel electrophoresis.
B) Three parallel ligations identical to that employed to construct biotinylated B+GC16 in part A were performed. Each was incubated with Dynabeads and washed as described above. The three identical Dynabead extractions of biotinylated indexed B+GC16 were added to three 20-µl aliquots of a master PCR reaction identical in composition to that used to amplify B+GC16 in part A. Each assembled reaction was amplified for 35 cycles using a different set of cycling parameters: "95-55-72" (30 s at 95°C; 30 s at 55°C; 60 s at 72°C; 35 cycles), "92-55-72" (30 s at 92°C; 30 s at 55°C; 60 s at 72°C; 35 cycles) or "92-55&59-72" (30 s at 92°C; 30 s at 55°C; 60 s at 72°C; 2 cycles; then 30 s at 92°C; 30 s at 59°C; 60 s at 72°C; 33 cycles) were employed.

FakePolyT fragments. Ligation, Dynabead extraction and PCR reaction assembly were performed in an identical manner to that employed in the previous experiment. Amplification was performed using three different PCR temperature cycling protocols. Analysis of the results (FIGURE 4.13B) indicated that amplification of the B^+GC16 template with Bam and GCRichPoly$(T)_{16}$-V primers was most efficient using the following PTC cycling parameters: a dissociation temperature of 92°C (to reduce the loss of activity by thermostable polymerase) for 30 sec; an initial annealing temperature of 55°C (to establish a high specific amplification coefficient for the targeted amplicon) for 30 sec; and an elongation temperature of 72°C for 60 sec. After the first two cycles, the annealing temperature was elevated to 59°C during each of the subsequent 33 cycles to provide higher stringency for primer annealing. (A 10-minute elongation step at 72°C at the end of 35 cycles was performed in all PCR protocols evaluated.)

4.3.8.4 *Estimation of indexed 3'-end cDNA template requirement for amplification with GCRichPoly$(T)_{16}$-V and BamCC*

A coarse evaluation of the amplification efficiency provided by cDNA indexing protocols developed through the use of FakePolyT fragments was performed using the A^+GC16 template. Indexed A^+GC16 was amplified in parallel preparative PCR reactions, pooled, purified and quantitated. An initial 6 ng/µl stock was serially diluted by factors of 10^{-2}, 10^{-4}, 10^{-6}, 10^{-8} and 10^{-10}. Six nanograms of A^+GC16 corresponds to 30 fmol, or roughly 20 billion A^+GC16 template molecules. One microlitre from each serial dilution of A^+GC16 was added to each of 5 identical PCR reactions. Amplification proceeded with Bam and GCRichPoly$(T)_{16}$-V primers using the PCR temperature cycling protocol "92-55&59-72". Analysis by agarose gel electrophoresis (FIGURE 4.14) revealed that amplification of indexed 3'-end cDNA fragments was efficient for as few as 10^2 templates, or less, present in PCR using cDNA indexing protocols.

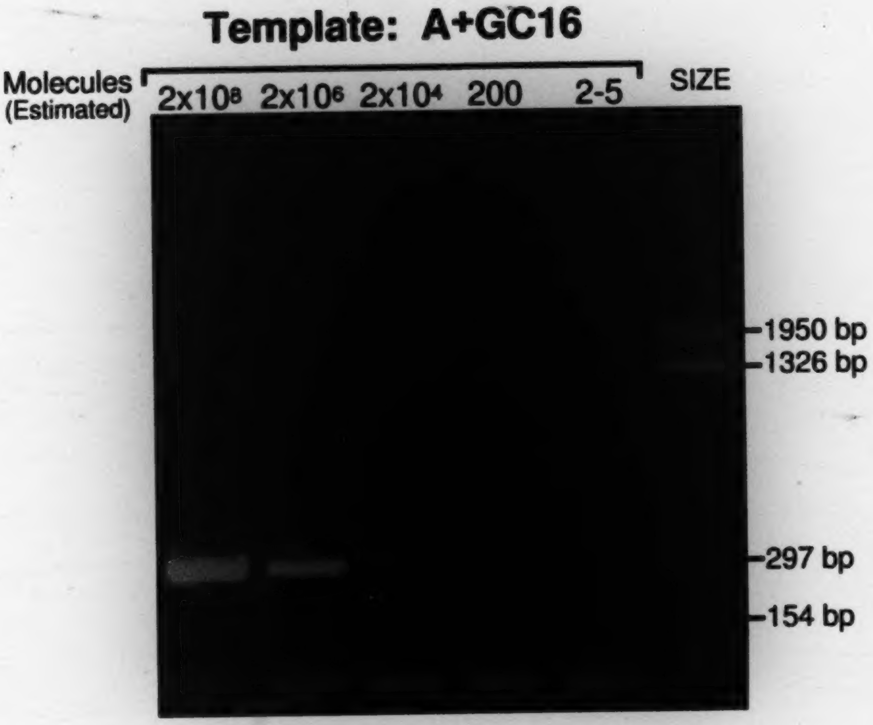


FIGURE 4.14: Coarse estimation of amplification efficiency for indexed GC16-tailed cDNA.
Indexed A$^+$GC16 was amplified in two parallel 50-µl PCR reactions from two 2-µl aliquots of the ligation employed in FIGURE 4.12. The entire pooled yield of the A$^+$GC16 amplification was isolated and purified by syringe extraction from agarose gel. These preparative amplifications yielded a total of approximately 90 ng of amplified A$^+$GC16 DNA. As A$^+$GC16 has a molecular mass of approximately 200 000, 90 ng of product corresponds to 450 fmol of A$^+$GC16. Resuspended purified template DNA was diluted to 30 fmol, or roughly 20 billion molecules, per microlitre. This stock was serially diluted by factors of 10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}, and 10^{-10}. One microlitre from each serial dilution was added to a 20-µl PCR reaction identical to that employed in FIGURE 4.13B. Amplification was performed using the PTC program "92-55&59-72", and PCR products were analyzed by agarose gel electrophoresis.

4.3.9 GC-rich anchored poly(T)-primed synthesis and quality assurance of ds-cDNA populations from total cellular mRNA isolated from *S. cerevisiae* cultures grown under different environmental conditions

The modified cDNA indexing protocols developed using Test Fragment primer pairs, biotinylated indexed pUC19 fragment combinations and FakePolyT fragments were applied to the generation and study of indexable ds-cDNA populations from *S. cerevisiae* mRNA. Cultures of *S. cerevisiae* strain W303 were grown in glucose medium and in galactose medium. Cultures growing exponentially in glucose medium either received no treatment, were treated with pheromone, or were exposed to high salt concentrations prior to the isolation of total cellular mRNA from those cells. mRNA populations isolated from these RNA preparations were used as the templates for ds-cDNA synthesis using a modified protocol (**Section 4.2.7**) and employing GCRichPoly(T)$_{16}$-V to initiate first-strand cDNA synthesis.

TF primer pairs were again used to evaluate the quality of cDNA synthesis for each ds-cDNA population. For each cDNA population, a set of four PCR reactions was performed using each of the four transcript-specific TF primer pairs. Analysis of PCR products by agarose gel electrophoresis (FIGURE 4.15) demonstrated amplification of each of the ACT1, FUS1, STE2 and STE12 TF transcripts from each of the ds-cDNA populations, indicating efficient and accurate synthesis of ds-cDNA from the isolated mRNA populations. Of particular note were the increased intensities of the product bands corresponding to FUS1 and STE12 TF amplification from the Pheromone ds-cDNA population, relative to the intensities of those bands for ds-cDNA populations derived from yeast grown under different growth conditions. Pheromone treatment of MATa *S. cerevisiae* cells induces transcription of the *STE12* gene coding for mating transcription factor and of the *FUS1* gene essential for fusion of MATa cells with cells of the opposite mating type [234, 237, 239, 240]. The increased prevalence of the mRNA transcripts of these two genes in the mRNA population derived from pheromone-treated yeast cells is reflected in the intensities of the bands produced by amplification to the FUS1 and STE12 TF fragments amplified from the corresponding cDNA population. This data supports the conclusion that the ds-cDNA populations generated using the modified cDNA synthesis protocol are of high quality and reflect the relative prevalence of the particular mRNA species of which the original mRNA populations were composed.

Each ds-cDNA population was digested with *Fok*I endonuclease in the manner outlined in **Section 4.2.5**. The quality of *Fok*I restriction digested of ds-cDNA populations was evaluated using TF primer pairs in a similar fashion to that described in **Section 4.3.2.2**. Agarose gel electrophoresis (FIGURE 4.15) demonstrated similar results for *Fok*I digestion of GC-rich poly(T)-primed cDNA populations as for digests of anchored poly$(T)_{16}$-primed cDNA populations.

4.3.10 Software development and database construction

The C^{++} program YeastORFdb v1.0 (developed by Eric Carpenter, Randy Nonay and Chris Dambrowitz) was developed to facilitate manipulation of *S. cerevisiae* open reading frame (ORF) cDNA transcript sequence data in preparation for 3'-end cDNA indexing analysis. The entire set of open reading frame sequences of *S. cerevisiae* (12,052 kb; 6,183 ORFs larger than 100 amino acids) was downloaded from the Saccharomyces Genome Database (SGD) via ftp as a compressed FASTA file [247] containing the DNA coding sequence for all current standard ORFs with introns removed. YeastORFdb opened the file and searched each individual ORF sequence for *Fok*I recognition site motifs in both the 5' and 3' direction. (A simple modification of the software permitted the recognition site of any Type IIS restriction endonuclease of interest to selected as the search motif.) In the case in which an ORF contained more than one *Fok*I recognition site, the program identified the site proximal to the 3'-end of the anticipated transcript. For each 3'-end fragment identified, the program generated a set of data entries describing the base position number of the start of the cutsite (counting from the final non-poly(A) base of the ORF sequence), the distance in bases from the cutsite to the end of the non-poly(A) region, and the sequence of the four-base cohesive end generated by *Fok*I cleavage from that recognition site of cDNA generated

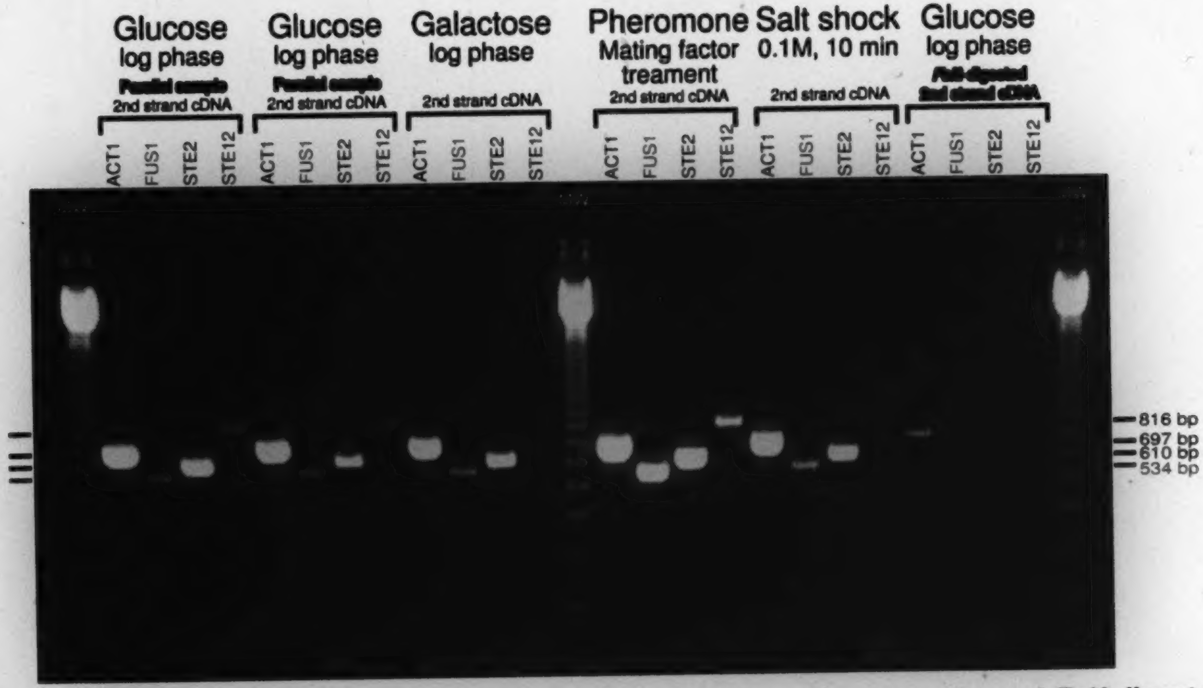

Glucose log phase
Parallel sample
2nd strand cDNA
Glucose log phase
Parallel sample
2nd strand cDNA
Galactose log phase
2nd strand cDNA
Pheromone Mating factor treament
2nd strand cDNA
Salt shock 0.1M, 10 min
2nd strand cDNA
Glucose log phase
FokI-digested 2nd strand cDNA
ACT1
FUS1
STE2
STE12
816 bp
697 bp
610 bp
534 bp

FIGURE 4.15: Evaluation of GC-rich anchored poly(T)-primed cDNA synthesis quality and *FokI* digest quality using TF primer pairs.
Cultures of *S. cerevisiae* strain W303 were grown in glucose medium and in galactose medium. Log-phase cultures growing in glucose medium received no treatment, were treated with pheromone, or were exposed to high salt concentrations prior to the isolation of total cellular mRNA from those cells. mRNA populations were isolated and used as templates for ds-cDNA synthesis. A modified protocol was employed (**Section 4.2.7**) using GCRichPoly(T)16-V for first-strand priming and utilizing the SuperScript kit for second-strand synthesis. For each cDNA population, a set of four PCR reactions was performed using each of the four transcript-specific TF primer pairs ACT1(red), FUS1 (green), STE2 (blue) and STE12 (purple), in order to evaluate the quality of cDNA synthesis. Each ds-cDNA population was digested with *FokI* endonuclease in the manner outlined in **Section 4.2.5**. The quality of *FokI* restriction digested of ds-cDNA populations was evaluated using TF primer pairs in a similar fashion to that described in FIGURE 4.3B (only Glucose *FokI*-digested cDNA shown). PCR reaction assembly, amplification conditions and analysis agarose gel electrophoresis were performed as described in FIGURE 4.3. Lanes marked "MW" contain 123-bp ladder molecular weight markers.

from that ORF's anticipated mRNA transcript. From these data, the program generated a second set of data entries corresponding to 3'-end cDNA fragments generated by *Fok*I restriction endonuclease: a fragment label corresponding to the ORF name and Locus name within the *Saccharomyces* genome; the base position numbers corresponding to the 3'-end of the fragment; the length of the indexed fragment in bases including the four-base cohesive end, indexer sequence and poly(T) tail; and the sequence of the indexer required to target the cohesive end sequences at the 5'-end of the 3'-end cDNA fragment. The data entries generated by the program were saved as comma-delineated text in a second file labeled "ORFdata.txt". This file could be opened by Microsoft Excel 98 to create a versatile searchable database of indexable 3'-end cDNA fragments in *Fok*I-digested cDNA populations derived from *S. cerevisiae* cultures grown under varying environmental conditions. This database was employed in the design of 3'-end cDNA indexing investigations and in the analysis of experimental results.

4.3.11 Differential expression of *GAL1* and of *BOP3* reported by 3'-end cDNA indexing

The *GAL1* gene of *S. cerevisiae* encodes galactokinase, the enzyme catalyzing the first step in galactose metabolism. The Gal1p gene product also relieves the inhibition of the Gal4p transcriptional activator by Gal80p in a similar manner to the action of Gal3p, to which it exhibits a high degree of amino acid sequence homology [248]. The abundance of *GAL1* mRNA has been reported to be 21.8-fold higher in cells utilizing galactose as a carbon source than in cells growing in glucose-containing medium [72, 249]. The high level of induction of *GAL1* in response to growth on galactose medium made it an attractive subject for the demonstration of differential expression reporting by 3'-end cDNA indexing. The YeastORFdb database was used to identify the size and cohesive end sequence of the 3'-terminal *Fok*I restriction fragment of the ds-cDNA synthesized from the *GAL1* mRNA transcript. From the database, the expected length of the indexed *GAL1* 3'-end cDNA fragment was

1382 bp, and the unindexed target bore the cohesive end sequence AACG. This cohesive end sequence would be targeted by the OH-CGTNxBamCC indexer mix.

The function of the protein encoded by the *BOP3* gene is unknown, but the expression of this gene is coregulated with a set of genes involved in the MAP kinase cascade inducing Ste12p and initiating filamentous growth in response to certain types of environmental stress [250]. A number of these genes (e.g. *AFR1*, *AGA1*, *CIK1*, *FAR1*, *FIG1*, *FUS1*, *KAR4*) are also upregulated in response to the mating signal transduction pathway [250]. Roberts *et al.* detected a 15-fold induction of *BOP3* in response to overexpression of *STE12* under control of the *GAL1* promoter in galactose-containing medium [73]. Under a variety of stress conditions, the expression of the *BOP3* mRNA transcript increases approximately 2-fold (Yeast Microarray Global Viewer [226, 230]). However, Causton *et al.* observed 3-fold repression of *BOP3* transcription following 15-minute treatment with 0.4 M NaCl [251]. The differential expression of *BOP3* by yeast grown in a variety of environmental conditions, in addition to the limited dynamic range of its expression, made this gene an attractive subject for differential expression reporting by 3'-end cDNA indexing. The YeastORFdb database indicated a 133-bp indexed amplicon was expected for the *BOP3* 3'-terminal restriction fragment. The GCTG cohesive end sequence of the *Fok*I-digested cDNA fragment would be targeted by the OH-CTGNxBamCC indexer mix.

A 3'-end cDNA indexing experiment was designed to detect and report differential expression of *GAL1* and of *BOP3* between a culture of *S. cerevisiae* strain W303 grown in galactose-containing medium and three cultures grown in glucose medium and subjected either to no treatment, exposure to α mating pheromone, or high-osmolarity shock (see **Section 4.2.1**). Four pairs of cDNA indexing ligations were assembled. Each pair of ligations contained 10 ng of *Fok*I-digested cDNA derived from yeast cultures grown under the various growth conditions. In each reaction, *Fok*I-digested cDNA was targeted by 50 fmol of either the OH-CGTNxBamCC or the OH-CTGNxBamCC biotinylated indexer mixes. The OH-CGTNxBamCC indexer mix targeted a set of 3'-terminal cDNA restriction fragments including the *GAL1* transcript. The OH-CTGNxBamCC indexer mix was used to target a set of 3'-terminal cDNA restriction fragments including the *BOP3* transcript. Ligation was performed using 40 U T4 DNA ligase at 37°C for 60 min, followed by heat denaturation of the ligase. Five micrograms of rinsed Dynabeads were added to each completed reaction, incubated and washed as before, and added to a PCR reaction containing BamCC and GCRichPoly(T)$_{16}$-V primers. Following 35 cycles of amplification using the PTC program "92-55&59-72", the PCR products were analyzed by agarose gel electrophoresis (FIGURE 4.16).

The set of cDNA transcripts detected in this experiment were identified by the YeastORFdb database and their corresponding gene functions ascertained by searching the YPD *Saccharomyces cerevisiae* Proteome Database [252, 253] (TABLE 4.3). The indexed *GAL1* 3'-terminal cDNA restriction fragment was detected in the cDNA population derived from yeast grown in galactose-containing medium, indicating a high level of expression of the *GAL1* mRNA transcript in those cells. The result obtained by expression reporting via cDNA indexing correlated with the known pattern of regulation of the *GAL1* gene [72, 249]. No indexed *GAL1* cDNA was detected among cDNA populations derived from cells grown in the presence of glucose. This finding was well-correlated with the results of DNA microarray studies of global gene expression in yeast grown on various carbon sources, in which cultures growing in rich glucose-containing medium were found to contain less than 0.1 molecule per cell of *GAL1* mRNA [254].

The indexed *BOP3* 3'-terminal restriction fragment was detected in the cDNA population derived from untreated cells growing in rich glucose medium. The indexed *BOP3* cDNA was also amplified from cDNA derived from pheromone-treated cells, generating a more intense band on agarose gel than that observed from the Glucose cDNA population. No *BOP3* expression was detected in cDNA derived from salt-shocked yeast, or from yeast grown in galactose medium. The results obtained by 3'-end cDNA indexing for differential expression of the *BOP3* gene were well-correlated with those obtained by DNA microarray analysis of differential gene expression [73, 250, 251].

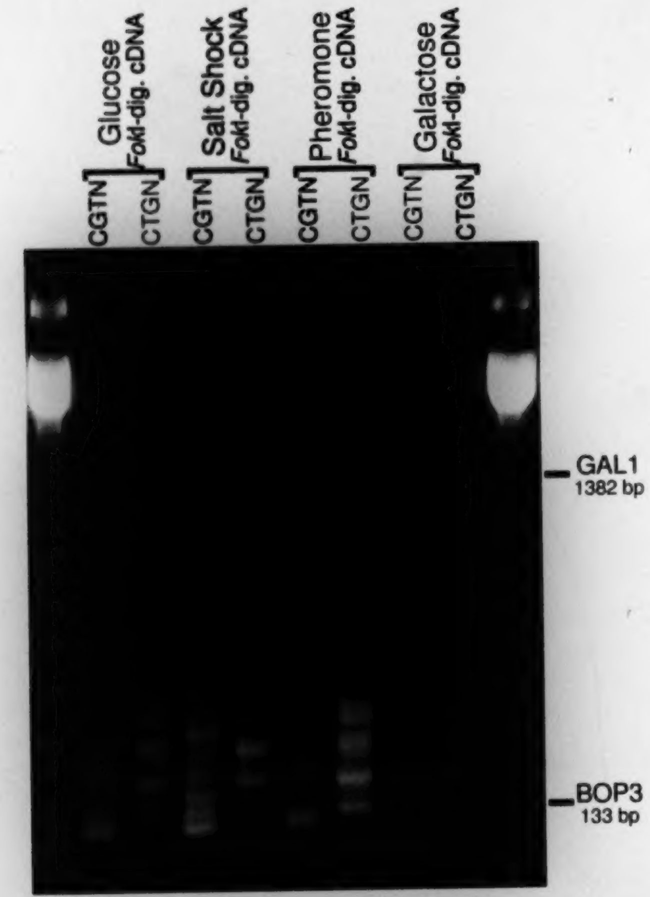

CGTN
CTGN
Glucose
FokI-dig. cDNA
CGTN
CTGN
Salt Shock
FokI-dig. cDNA
CGTN
CTGN
Pheromone
FokI-dig. cDNA
CGTN
CTGN
Galactose
FokI-dig. cDNA
GAL1
1382 bp
BOP3
133 bp

FIGURE 4.16: Differential expression of *GAL1* and of *BOP3* reported by 3'-end cDNA indexing.

Four pairs of cDNA indexing ligations were assembled to demonstrate differential expression of *GAL1* and of *BOP3*. Each pair of ligations contained 10 ng of *Fok*I-digested cDNA derived from yeast cultures grown under various growth conditions. In each reaction, *Fok*I-digested cDNA was targeted by 50 fmol of either the OH-CGTNxBamCC or the OH-CTGNxBamCC biotinylated indexer mixes. The OH-CGTNxBamCC indexer mix targeted a set of 3'-terminal cDNA restriction fragments including the *GAL1* transcript. The OH-CTGNxBamCC indexer mix was used to target a set of 3'-terminal cDNA restriction fragments including the *BOP3* transcript. Ligation was performed using 40 U T4 DNA ligase at 37°C for 60 min, followed by heat denaturation of the ligase. Five micrograms of rinsed Dynabeads were added to each completed reaction, incubated and washed as before, and added to a PCR reaction containing BamCC and GCRichPoly(T)16-V primers. Following 35 cycles of amplification using the PTC program "92-55&59-72", the PCR products were analyzed by agarose gel electrophoresis. Lanes marked "MW" contain 123-bp ladder molecular weight markers.

TABLE 4.3: Identification of differentially-expressed transcripts by 3'-end cDNA indexing.

INDEXER		GROWTH CONDITION				FUNCTION
Gene	Size	GLU	SALT	PHER	GAL	
CGTN						
YCR105W	101	Y	Y	Y	Y	NADPH-dependent cinnamyl-alcohol dehydrogenase
SEH1	115	Y	Y	W	W	protein in complex with nuclear pore protein
SPO69	125	N	Y	N	N	required for sporulation and meiosis
RPL17A	188	Y	Y	Y	Y	basic protein in 60S ribosomal subunit
YBL010C	268	N	Y	N	N	proteins of unknown function
YBR007C	269					
YNR042W	269					
YKR051W	452	Y	W	W	Y	protein of unknown function
GAL1	1382	N	N	N	Y	galactokinase
CTGN						
BTN2	109	Y	Y	Y	N	involved in cellular pH homeostasis
BOP3	133	Y	N	Y	N	overproduction inhibits growth and induces filamentous phenotype
YNL247W	187	Y	Y	Y	Y	cysteinyl-tRNA synthetase
YDR262W	255	Y	Y	Y	Y	putative transmembrane protein
DOT5	302	Y	Y	Y	Y	chromatin structure and cell stress
YNL081C	329	Y	N	Y	N	putative mitochondrial ribosomal protein
ISD2	330					mitochondrial protein required for iron metabolism
DIB1	430	Y	N	W	N	U4/U6.U5 snRNP component with role in mitotic spindle formation

Amplified indexed 3'-end cDNA transcripts detected in FIGURE 4.16 were identified using the YeastORFdb cDNA indexing database. Amplified 3'-end cDNA transcripts that could not be uniquely identified from the Yeast ORFdb database due to limitations of agarose gel electophoretic separation were included using each potential transcript identity, and are highlighted in grey. Function of the corresponding gene (if known) was identified from YPD [252].

Significant background amplification was evident among the PCR products generated in this experiment. In subsequent protocol modification, efforts were made to establish ligation conditions which reduced the level of background amplification in 3'-end cDNA indexing reactions.

4.3.12 Conditions providing high ligation fidelity by *Taq* DNA ligase

An extensive investigation into ligation fidelity in cDNA indexing reactions and methods by which it might be improved was recently described by Shaw-Smith *et al.* [223]. These authors identified a set of conditions under which *Taq* DNA ligase could be used effectively to provide significantly enhanced ligation specificity over T4 DNA ligase. Under the conditions defined by this study, ligations containing a molar ratio of indexer to cDNA cohesive end targets of 10:1 with 40 U *Taq* DNA ligase incubated at 14°C for 60 minute provided a ligation specificity approaching 100%. The authors applied this modified cDNA indexing ligation protocol to the identification of differences in gene expression in murine duodenum and ileum [219]. The conditions developed by Shaw-Smith *et al.* were modified and incorporated with the streptavidin-capture and amplification protocols developed for 3'-end cDNA indexing investigations of global gene expression in *S. cerevisiae*.

4.3.13 Reporting by 3'-end cDNA indexing of differential gene expression in yeast in response to saline stress

The challenge presented to *Saccharomyces* by conditions of high salinity evokes a range of physiological responses including measures to ameliorate osmotic stress and to counter the toxicity of specific cations [255, 256]. A complex transcriptional program involving large numbers of genes is induced that activates these and other biochemical functions [256-258]. Analysis of the global transcriptional response of the yeast genome under NaCl stress has been performed using DNA microarrays [257, 259].

The 3'-end cDNA indexing protocol set incorporating ligation conditions modified from Shaw-Smith *et al.* [223] was initially applied to the generation of partial gene expression profiles for *S. cerevisiae* grown in glucose-containing medium with or without short-term exposure to high salt concentrations (**Section 4.2.1**). A set of seven ligations using indexer mix sequences AACN, CGCN, CGTN, CTCN, CTGN, GACN and GGGN was performed for the Glucose cDNA population and a set for the Salt cDNA population. Each 3'-end cDNA indexing ligation reaction contained 1 ng of *Fok*I-digested ds-cDNA derived from yeast grown in glucose medium with or without saline treatment, 500 fmol (per indexer) of the appropriate biotinylated nonphosphorylated BamCC indexer mix and 10 U *Taq* DNA ligase in a 15-µl reaction volume of *Taq* DNA ligase buffer. Ligations were incubated at 16°C for 2 h. Fifteen microlitres of a Dynabead suspension in 2 M NaCl was added to each cDNA indexing ligation containing biotinylated cDNA fragments. The samples were incubated at 37°C for 1 h with gentle agitation, washed using a stringent wash regimen, rinsed and the liquid removed. A 18-µl volume of Platinum™ *Taq* DNA polymerase reaction mix was prepared with 40 pmol BamCC indexing primer; 60 pmol [total oligo] GCRichPolyT$_{16}$-V (20 pmol/anchored end) and 1.25 U Platinum™ *Taq* DNA polymerase. This PCR reaction mix was added onto the magnetically-isolated Dynabeads to which the biotinylated cDNA fragments were bound, and mixed thoroughly by pipetting up and down. Terminal cDNA targets bearing a biotinylated BamCC indexer on one end and a GCRichPoly(A) region at the other were amplified using the PTC protocol "92-55&59-72" for 35 cycles. To prepare samples for analysis by agarose gel electrophoresis, 4 µl of loading dye was added directly to the 20-µl PCR reaction and mixed thoroughly. Twelve microlitres of each dyed reaction was loaded onto a 2% agarose gel and electrophoresed for 120 min at 8 V/cm. Appropriate DNA size standards were included on the gel. Visualization of electrophoresis results was performed by UV transillumination as previously described, and the data were recorded by digital image capture (FIGURE 4.17).

Identification of amplified indexed 3'-end cDNA transcripts was performed using the YeastORFdb cDNA indexing database (TABLE 4.4). For each transcript

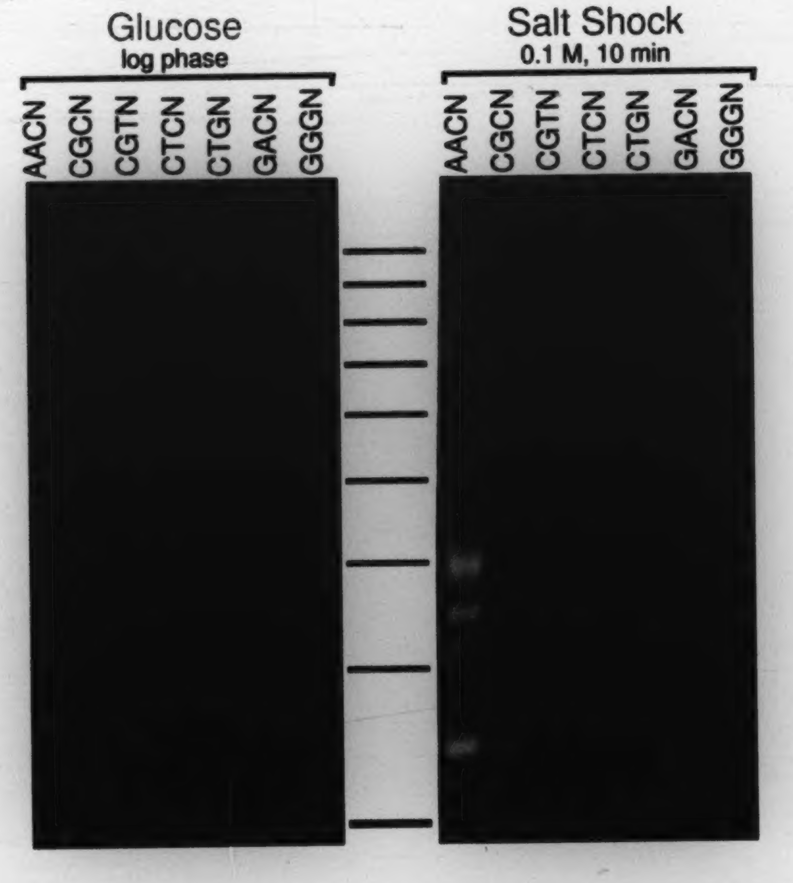

Glucose
log phase
AACN
CGCN
CGTN
CTCN
CTGN
GACN
GGGN
Salt Shock
0.1 M, 10 min
AACN
CGCN
CGTN
CTCN
CTGN
GACN
GGGN

FIGURE 4.17: Differential gene expression in yeast in response to saline stress.
Two sets of seven ligations using indexer mix sequences AACN, CGCN, CGTN, CTCN, CTGN, GACN and GGGN were performed for the Glucose and Salt cDNA populations. Ligations were performed using *Taq* DNA ligase (see text) and incubated with Dynabeads for streptavidin capture of biotinylated DNA fragments. PCR was performed on the Dynabead-bound DNA templates using BamCC GCrichPolyT$_{16}$-V primers. Analysis of 3'-end cDNA indexing results was performed using agarose gel electophoresis. Amplified indexed 3'-end cDNA transcripts were identified using YeastORFdb (TABLE 4.4).

TABLE 4.4: Identification of indexed cDNA transcripts expressed in response to saline stress.

INDEXER		GROWTH CONDITION		INDUCTION*	FUNCTION
Gene	Size	Glucose	Salt		
AACN					
FPR4	183	Y	Y	2.8	nucleolar peptidyprolyl cis-trans isomerase (PPIase)
GCD2	187			1.8	translation initiation factor, 71 kDA subunit
DPS1	224	N	Y	6.5	cytoplasmic aspartatyl-tRNA synthetase
YDR526C	269	Y	Y	2	unknown function; null mutant is lethal
SPT5	296	Y	Y	2.6	role in chromatin structure; influences expression of many genes
YDR426C	370	Y	Y	1.7	protein of unknown function
DED81	499	N	Y	9.3	cytoplasmic asparaginyl-tRNA synthetase
CGCN					
DAK2	172	Y	Y	1.2	dihydroxyacetone kinase
SAN1	664	N	Y	3.4	role in chromatin structure
CGTN					
YDR107C	248	Y	Y	1.5	role in small molecular transport
ISA1	258	Y	Y	1	mitochondrial protein required for normal iron metabolism
NNF1	260			0.8	nuclear envelope protein required for nuclear migration
YKR051W	452	Y	N	1	protein of unknown function
MDL1	505	Y	N	1.2	member of ATP-binding cassette superfamily
SAS10	505			1.2	role in chromatin structure
YFR020W	590	Y	Y	1.3	protein of unknown function
CTCN					
ROM1	258	Y	N	1.5	signal transduction and cell wall maintenance
YMR31	327	Y	Y	1.1	mitochondrial ribosomal protein
STH1	327			1.2	component of chromatin remodeling complex
UBP2	481	Y	N	1.6	ubiquitin-specific protease
AEP2	571	Y	N	1.2	mitochondrial protein
HST4	625	Y	N	0.7	histone deacetylase; role in transcriptional silencing of chromatin
YNL191W	870	Y	N	1	protein of unknown function
CTGN					
YMR222C	255	N	Y	3	protein similar to *S. pombe* dihydrofolate reductase
SKO1	322	Y	Y	2.1	suppressor of protein kinase A overexpression
RPB8	323			3.2	shared subunit of RNA polymerases I, II and III
HIS3	325			1.5	involved in histidine biosynthesis
ADE4	451	N	Y	1.3	catalyzes first step in *de novo* purine biosynthesis
GACN					
OM45	192	Y	Y	3.4	protein of the outer mitochondrial membrane
CKI1	361	Y	Y	2.9	choline kinase
VAM7	439	Y	Y	1.7	subunit of vacuolar SNARE complexes
YKL214C	517	Y	Y	1.5	protein associated with RNP complexes
CNA1	547	Y	N	0.7	protein serive/threonine phosphatase 2B
ZAP1	550			1.3	zinc-responsive transcriptional activator
YCL010C	799	N	Y	1.3	protein of unknown function
GGGN					
SAP1	220	Y	Y	1	role in chromatin/chromosome structure
YHR176W	220			1.1	catalyst of biological thiol oxidation
PPG1	549	N	Y	2.3	serine/threonine phosphatase involved in glycogen accumulation

* Levels of induction of particular mRNAs following saline treatment of *S. cerevisiae*, as observed through cDNA microarray analysis by Popas *et al.* [257]. Induction greater than1.5-fold is noted in red; repression is noted in blue. Indexed 3'-end cDNA transcripts not uniquely identifiable from the Yeast ORFdb database are highlighted in grey. Function of the corresponding gene (if known) was identified from YPD [252].

detected for either the Glucose or Salt cDNA population, the function of the corresponding gene (if known) was established by searching YPD [253]. The levels of induction of particular mRNAs following saline treatment of *S. cerevisiae*, as observed through DNA microarray analysis by Popas *et al.* [257] were identified for comparison. Amplified 3'-end cDNA transcripts that could not be uniquely identified from the YeastORFdb database due to limitations of agarose gel electophoretic separation were incorporated in the Table using each potential transcript identity.

This experiment demonstrated the successful application of the set of modified 3'-end cDNA indexing protocols developed during this investigation. In particular, ligations performed using *Taq* DNA ligase according to a protocol modified from that described by Shaw-Smith *et al.* [223] were efficient and generated low background amplification compared to previous experiments performed using T4 DNA ligase. Differential gene expression between saline-treated and untreated yeast cultures was observed by 3'-end cDNA indexing.

A previous investigation of global transcriptional changes induced in *Saccharomyces* in response to high salinity had been performed using DNA microarray analysis [257]. A different yeast strain, *S. cerevisiae* strain TM141 (*MAT*a *ura3 leu2 trp1 his3*), was used as the basis for that investigation. Additionally, the researchers had used similar but nonidentical saline-challenge conditions to those employed in this 3'-cDNA indexing experiment (10 min exposure and 20 min exposure to 0.4 M NaCl, compared to 15 min exposure to 0.4 M NaCl for this study). They identified substantial differences between the two transcriptional programs induced in strain TM141 for these two saline-exposure periods, indicating a very transient expression pattern across a large number of genes. It has been recognized that seemingly subtle differences in strain genetics or in culture conditions in DNA microarray investigations of global gene expression can result in substantially different transcription profiles [260, 261]. Even data sets generated using nominally identical experimental conditions and the same commercial oligonucleotide array have been reported to be poorly correlated [184, 186, 262]. These differences have been attributed to low precision in mRNA concentration estimates obtained from microarrays [213, 214] and to minor strain, preparation and growth condition differences [262]. Despite the differences in the genetic background of strain TM141 investigated using DNA microarrays and strain W303 employed in this cDNA indexing experiment, and despite the small but important differences in experimental procedure between the two investigations, the gene expression profiles generated in this experiment by 3'-end cDNA indexing were largely compatible with data obtained by DNA microarray. The limitations of agarose gel electrophoresis as a method of data analysis for 3'-end cDNA indexing (see below) did not prevent the generation of informative data during this demonstration of cDNA indexing protocols.

While agarose gel electrophoresis was a useful and simple method of data analysis during protocol development for 3'-end cDNA indexing of *S. cerevisiae*, the limitations of this separation technique for full exploitation of the information content of global gene expression profiling experiments were apparent. Size determination of amplification products to single-base resolution was not possible by this technique. The relatively low resolution of amplification products was insufficient to permit unique identification of several 3'-end cDNA transcripts from the database. The sensitivity of amplicon detection provided by UV transillumination of EtBr-stained agarose gels was low, making detection of weakly-amplified amplicons of transcripts present at low copy number in PCR difficult. Attempts to provide approximate quantitation were complicated by differential detection of larger fragments as a result of increased EtBr staining intensity. Taken together, these limitations reduce the utility of agarose gel electrophoresis as a method of data analysis for 3'-end cDNA applications, despite its utility during protocol development. Subsequent analysis of gene expression profiles generated by 3'-end cDNA indexing was performed using automated fluorescence-based DNA sequencing instrumentation.

4.3.14 3'-end cDNA indexing data acquisition and analysis by automated DNA sequencing instrumentation

The fragment size resolution provided by agarose gel electrophoresis, and the sensitivity of detection of cDNA indexing amplification products by UV transillumination of EtBr-stained agarose gels, were inadequate to fully exploit the information content of 3'-end cDNA indexing studies of yeast transcription profiles. In subsequent global gene expression profiling studies, fluorescently-labeled cDNA indexing samples were analyzed using automated DNA sequencing instrumentation. In cDNA indexing experiments to be analyzed by this approach, ligation with a set of NoP indexer mixes and subsequent streptavidin capture of biotinylated indexed 3'-terminal cDNA fragments were performed as in the previous experiment. BamCC indexing primers labeled with one of the fluorescent dyes FAM, JOE, or ROX were used in PCR. Prior to analysis, 3'-end cDNA amplifications indexed with identical nonphosphorylated BamCC indexer mixes but originating from different yeast cDNA populations (and thus labeled with different fluorescent dyes) were pooled. The pooled samples were purified with the Concert™ Rapid PCR Purification System and concentrated into 5 to 10 µl with Microcon YM-30 spin columns. A TAMRA-labeled DNA size standard (LargeFrag DNA size standard) was generated by amplifying 666-bp, 826-bp and 983-bp fragments from a duck hepatitis B virus (D-HBV) size standard set with a TAMRA-labeled common forward primer [232], and the reactions were pooled and concentrated in a similar manner to the indexed cDNA fragments. Two microlitres of GeneScan-500 [TAMRA] DNA size standard (Applied Biosystems) and 2 µl of the LargeFrag DNA size standard were added to each concentrated reaction and mixed thoroughly. The standard-doped indexed cDNA reactions were stored at -20°C in the dark until analyzed.

Data acquisition for fluorescently-labeled 3'-end cDNA fragment sets doped with TAMRA-labeled DNA standards was performed on an ABI 377 DNA Sequencer (Applied Biosystems) according to the manufacturer's instructions for RFLP genetic analysis. Experimental data sets were analyzed and manipulated using ABI GeneScan v3.1 genetic analysis software for automated sequencers.

4.3.15 Evaluation of global gene expression profiling in *S. cerevisiae* by 3'-end cDNA indexing

In order to evaluate the modified 3'-end cDNA indexing protocols for yeast transcriptomics using automated DNA sequencing instrumentation, global gene expression profiles generated from saline-treated yeast were compared to profiles generated from yeast growing logarithmically in glucose medium without saline treatment. Four sets of 32 cDNA indexing ligations were assembled. Each ligation in a set contained one of the first 32 BamCC NoP indexer mixes (bearing cohesive end sequences AAAN to CTTN). Set 1 targeted the Salt *Fok*I-digested cDNA population (generated from yeast grown logarithmically in glucose medium, and subjected to high-osmolarity salt shock with 0.4 M NaCl for 15 min). Set 2 was a replicate set of ligations also targeting the same Salt *Fok*I-digested cDNA population. (This would permit evaluation of the reproducibility of cDNA indexing expression profiles generated from identical cDNA populations.) Set 3 and Set 4 each targeted one of two separate *Fok*I-digested cDNA populations (Glucose and Glucose*, respectively) derived from independent parallel yeast cultures grown logarithmically in glucose medium without other treatment. Comparison of these two cDNA populations derived from independent cultures under identical conditions would allow evaluation of the reproducibility of cDNA population synthesis by the method employed in this investigation. In addition, comparison of the Glucose and Glucose* cDNA indexing profiles would indicate the level to which these profiles provided meaningful information regarding the biological significance of any similarities or differences noted between cDNA populations derived from dissimilar sources.

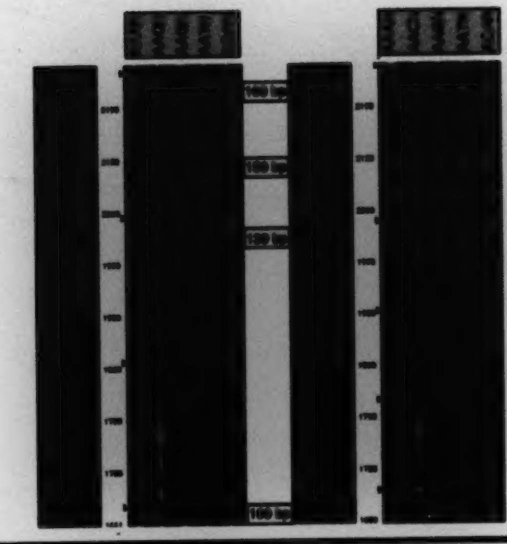
Following ligation and streptavidin capture, each set of 32 indexed samples were amplified using anchored GC-rich poly(T) and dye-labeled BamCC primers. Sets 1 and 2, the parallel Salt cDNA indexing samples, were amplified using FAM-labeled BamCC (corresponding to the "blue" channel of automated DNA analytical instrumentation). Sets 3 and 4, the Glucose and Glucose* sets, were amplified using JOE-labeled BamCC (corresponding to the "green" channel of the instrumentation used for analysis). Reactions from Sets 1 and 3 (Salt1 and Glucose) were pooled, purified and spiked with TAMRA-labeled DNA size standard. Reactions from Sets 2 and 4 (Salt2 and Glucose*) were similarly prepared for analysis. Each pooled reactions series of 32 samples was analyzed on an ABI 377 DNA sequencer according to the manufacturer's instruction, and the data analyzed as described above.

4.3.15.1 Reproducibility of 3'-end cDNA indexing profiles of gene expression in S. cerevisiae

In order to evaluate the reproducibility of amplicon patterns produced by 3'-end cDNA indexing, the profiles generated from parallel indexing iterations of the *Fok*I-digested Salt cDNA population were compared. A detailed comparison of a small subset of the cDNA indexing profiles generated from parallel Salt1 and Salt2 indexing sets is illustrated in FIGURE 4.18. The cDNA indexing products between 100 bp and 160 bp in length targeted by indexer sequences AAAN, AACN, AAGN, AATN and ACAN are shown for the Salt1 and Salt2 cDNA indexing sets (FIGURE 4.18A). The patterns of amplification products generated from independent indexing ligations targeting the same complex population of *Fok*I-digested cDNA fragments were reproducible. Amplification products corresponding to particular transcripts identified from the database are described in the Table (FIGURE 4.18B). Despite the significant number of indexing amplicons which correlated to identifiable cDNA transcripts, a surprisingly large number of unanticipated amplicons were also detected using automated fluorescence-based instrumentation and software. These unanticipated amplicons present a significant challenge to the utility of 3'-end cDNA indexing in *S. cerevisiae*.

Similarly, the reproducibility of cDNA indexing profiles generated from distinct cDNA populations derived from similar cultures grown under identical conditions was evaluated by comparison of the indexing profiles produced from the Glucose and Glucose* *Fok*I-digested cDNA populations. FIGURE 4.19 demonstrates the comparison of a small subset of the Glucose and Glucose* cDNA indexing profiles, illustrating the indexing amplicons detected between 100 bp and 160 bp in reactions targeted by indexer sequences CGGN, CGTN, CTAN and CTCN. As in the previous

A



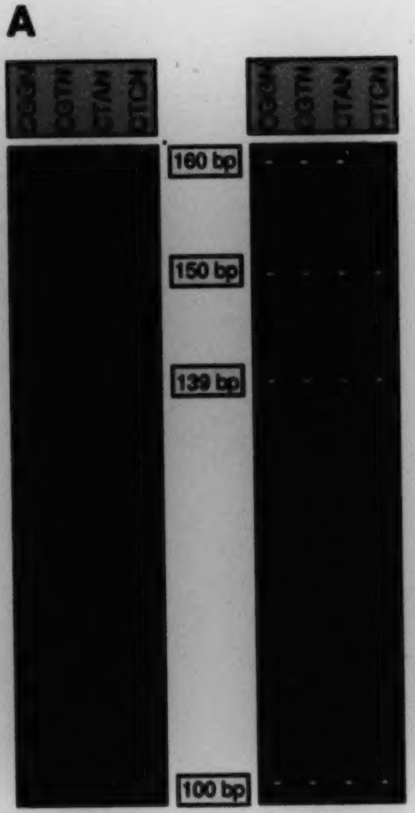
B

ORF	Protein Function	Indexer	Fragment Length (bp)
PIP2	protein involved in induction of peroxisomal proteins in response to oleic acid	AAAN	100
FKH1	forkhead protein homolog with roles in transcriptional silencing and cell cycle	AAAN	113
MDJ1	chaperone protein involved in mitochondrial biogenesis and protein folding	AAAN	115
YCL069W	member of multidrug resistance 14 spanner family	AAAN	128
YKR105C	member of multidrug resistance 14 spanner family	AAAN	128
PDR10	member of ATP-binding cassette superfamily	AAAN	143
SQT1	protein required of ribosomal assembly	AAAN	153
YOR309C	protein of unknown function	AAAN	160
YHR210C	protein of unknown function, similar to aldose epimerases	AACN	108
MCM6	protein involved in DNA replication, member of MCM/P1 superfamily	AACN	111
RIM13	sporulation protein involved in proteolytic processing of Rim101p	AACN	112
MED8	component of RNA polymerase II holoenzyme and mediator subcomplex	AACN	112
KRE9	glycoprotein involved in cell surface β-1,6-glucan assembly	AACN	113
ARR2	arsenate reductase	AACN	136
YCR051W	protein with similarity to ankytin and coiled-coil protein	AACN	148
YMR122C	protein of unknown function	AAGN	108
LST7	protein needed for tranport of N-regulated permeases from Golgi to plasma membrane	AAGN	118
YMR010W	protein of unknown function	AAGN	133
PMT1	mannosyl transferase	AAGN	157
SPT21	protein amplifies magnitude of transcriptional regulation at various loci	AATN	133
VMA8	vacuolar H+-ATPase	AATN	154
YOR331C	protein of unknown function	ACAN	116
HTA1	histone H2A	ACAN	124
YGR010W	possible nicotinamide mononucleotide adenylyltransferase	ACAN	130
SIN4	component of RNA polymerase II holoenzyme and mediator subcomplex	ACAN	139

FIGURE 4.18: Reproducibility of 3'-end cDNA indexing profiles of gene expression in *S. cerevisiae* 1: parallel indexing iterations of an individual *Fok*I-digested cDNA population.

A) A detailed comparison of a subset of the cDNA indexing profiles generated from parallel Salt1 and Salt2 indexing sets was made to evaluate the reproducibility of amplicon patterns produced by 3'-end cDNA indexing,. cDNA indexing products between 100 bp and 160 bp in length targeted by indexer sequences AAAN, AACN, AAGN, AATN and ACAN are shown for the Salt1 and Salt2 cDNA indexing sets. See Section 4.3.15 for further experimental details.

B) Amplified indexed 3'-end cDNA transcripts were identified using the YeastORFdb cDNA indexing database. For each transcript detected in the Salt1 and Salt2 cDNA indexing profiles, the function of the corresponding gene (if known) was established by searching YPD [253].



B

ORF	Protein Function	Indexer	Fragment Length (bp)
SUR4	required for conversion of 24-carbon fatty acids to 26-carbon fatty acids	CGGN	122
GRD19	involved in retrieval of Golgi membrane proteins from prevacuolar complex	CGGN	122
YLR064W	protein of unknown function	CGGN	133
YJR084W	protein of unknown function	CGGN	154
YCR105W	NADPH-dependent cinnanyl-alcohol dehydrogenase	CGTN	101
YMR318C	NADPH-dependent cinnanyl-alcohol dehydrogenase	CGTN	101
SEH1	protein found in complex with nuclear pore proteins	CGTN	115
SPO69	protein required for meiosis and sporulation	CGTN	125
PRI2	DNA primase subunit	CGTN	153
HCA4	probable RNA helicase CAC1	CGTN	155
HPA2	histone acetyltransferase	CTAN	121
STT3	oligosaccharyltransferase subunit	CTAN	139
ECM22	protein with similarity to transcription factors	CTAN	139
MRS6	protein required for protein transport from ER to Golgi apparatus	CTAN	151
YDR287W	inositol monophosphatase	CTCN	108
ESR1	checkpoint protein kinase required for mitotic growth, DNA repair and recombination	CTCN	142
FUS1	protein required for cell fusion during mating, localized to shmoo projection	CTCN	159
YKR018C	protein of unknown function	CTCN	160

FIGURE 4.19: Reproducibility of 3'-end cDNA indexing profiles of gene expression in *S. cerevisiae* 2: parallel indexing iterations of parallel *FokI*-digested cDNA populations.
A) A comparison of a small subset of the Glucose and Glucose* cDNA indexing profiles was made, illustrating the indexing amplicons detected between 100 bp and 160 bp in reactions targeted by indexer sequences CGGN, CGTN, CTAN and CTCN. See Section 4.3.15 for further experimental details.
B) Identification of amplified indexed 3'-end cDNA transcripts was performed using the YeastORFdb cDNA indexing database. For each transcript detected in the Glucose and Glucose* cDNA indexing profiles, the function of the corresponding gene (if known) was established by searching the Yeast Proteome Database [253].

figure, amplicons corresponding to particular transcripts identified from the database are described in the Table (FIGURE 4.19B). The amplification profiles generated from the Glucose and Glucose* cDNA populations demonstrated adequate reproducibility. However, as seen in the comparison of parallel indexing of the Salt cDNA population, a large number of unanticipated fragments were amplified to detectable levels from the Glucose and Glucose* cDNA populations. Also, the level of background amplification generated in both sets of cDNA indexing reactions was disappointingly high, particularly in the Glucose* reaction set, and may have obscured the detection and identification of legitimate indexed amplicons present at low levels.

4.3.15.2 Comparison of cDNA indexing gene expression profiles generated from saline-treated and untreated yeast cultures

To provide an indication of the extent to which the data obtained from 3'-end cDNA indexing of *Fok*I-digested cDNA populations reflected biologically significant changes in the global gene expression patterns in yeast in response to environmental stimuli, cDNA indexing data profiling the response of *S. cerevisiae* to high osmolarity challenge was compared to data obtained by cDNA microarray for similar experiments [257]. The authors of the cDNA microarray study identified a set of 121 yeast genes demonstrating a 5-fold or greater change in expression in response to exposure of the cells to 0.4 M NaCl for 10 minutes. The cDNA microarray data also demonstrated that the global gene expression pattern induced by this saline treatment was transient and that mRNA levels of most genes induced by 10-minute exposure to saline dropped dramatically following 20 minutes of exposure [257]. Thirty-eight of the 121 highly-saline-responsive ORFs were anticipated to be represented by 3'-terminal cDNA *Fok*I fragments targeted by the cohesive end sequences present in the 32 NoP indexer mixes OH-AAANxBamCC to OH-CTTNxBamCC. Of this subset, 28 fragments were expected to be between 100 bp and 550 bp in length, thus falling within the fragment size range most accurately characterized using the GeneScan-500 DNA size standard on automated DNA analytical instrumentation.

The 3'-end cDNA indexing profiles of yeast gene expression generated from the Salt and Glucose cDNA populations were analyzed and compared. Of the subset of 28 anticipated 3'-terminal cDNA fragments with the characteristics described above corresponding to saline-responsive ORFs identified by Posas *et al.*, 27 were detected in this experiment (TABLE 4.5). In an attempt to evaluate the extent to which the fluorescent intensity of a particular amplicon was indicative of the relative expression level of the mRNA transcript from which it was derived, the electropherogram data for peaks corresponding to (FAM-labeled) Salt and (JOE-labeled) Glucose indexed cDNA fragments were compared. The ratio of the electropherogram peak areas obtained for a particular indexed cDNA amplicon in the Salt and Glucose cDNA populations was determined for each 3'-terminal cDNA fragment corresponding to one of the 27 detected saline-responsive ORFs. Ratios reflecting less than two-fold differences in fluorescent peak intensity were not considered to be significant, in keeping with accepted practice for the interpretation of gene expression data obtained by cDNA microarray [230, 257, 263]. The Salt:Glucose peak area ratios for twenty of the 3'-end cDNA fragments indicated a two-fold or greater difference in the amount of amplicon present between the two cDNA sources. Although the peak area ratios of 17 of these cDNA fragments indicated an increased level of amplification of the target from the Salt cDNA population relative to the Glucose cDNA population, suggesting a relative increase in the initial copy number of indexed 3'-terminal cDNA fragments present prior to PCR corresponding to increased mRNA expression from the related ORFs in response to saline treatment, the magnitude of the Salt:Glucose peak area ratios is not well-correlated to the levels of induction reported using cDNA microarrays (TABLE 4.5). In some cases, the results obtained by 3'-end cDNA indexing were contradictory to the cDNA microarray data (e.g. *RPL37B, NTH1, STL1*).

Differences in gene expression profiles between the cDNA indexing and cDNA microarray data sets were anticipated due to the different durations of saline treatment performed in the two experiments, and given the transient nature of the saline-response program of gene expression observed by Posas *et al.* [257]. Slight differences in the genetic background of the yeast strains employed in the two investigations may have

TABLE 4.5: Reporting of differential gene expression in response to saline stress by 3'-end cDNA indexing.

Protein Function	ORF	Indexer	Fragment Length (bp)	Peak Area Ratio Salt:Glucose	Induction (from Posas *et al.* [257])	
					0.4 M NaCl, 10 min	0.4 M NaCl, 20 min
3-phosphoserine transaminase	SER1	AAAN	286	no significant change	17.6	1.8
Ribosomal protein S22	RPS22A	AAAN	368	no significant change	5.1	1.5
Vacuolar H^+-ATPase 94 kDa subunit	VPH1	AAAN	392	no significant change	11	1.8
Plasma membrane Na^+-ATPase pump	ENA1	AAAN	423	4.0	7.4	1.2
Subunit of vacuolar H^+-ATPase	VMA5	AAAN	532	24.0	8.9	2.0
Aspartyl-tRNA synthetase (cytosolic)	DPS1	AACN	224	56.4	6.5	1.2
Asparaginyl-tRNA synthetase (cytosolic)	DED81	AACN	499	not detected	9.3	1.3
Ribosomal protein YL37	RPL37B	AAGN	209	0.5	6.3	1.3
Branching enzyme - glycogen metabolism	GLC3	AATN	279	26.2	5.1	1.5
Neutral trehalase	NTH1	ACTN	171	0.1	12.0	2.6
Osmotic stress response protein	GRE2	AGGN	244	5.0	22.0	7.5
High-affinity hexose transporter	HXT7	ATGN	146	3.2	9.6	1.6
Ribosomal protein L21	RPL12A&B	ATTN	214	no significant change	5.9	1.2
Glutaminyl-tRNA synthetase	GLN4	ATTN	339	no significant change	5.4	1.0
Catalytic subunit of vacuolar H^+-ATPase	TFP1	ATTN	340	no significant change	6.3	1.4
Plasma membrane sugar transporter	STL1	CATN	265	0.2	89.8	2.2
Ribosomal protein L28	RPL28	CATN	366	9.8	5.9	1.9
Ribosomal protein S18	RPS11A&B	CCAN	337	13.5	5.1	1.0
Mitochondrial ribosomal protein	MRPL40	CCAN	472	25.1	8.8	2.2
Glycogen synthetase	GSY2	CCGN	247	2.0	24.8	3.1
Stress response protein of unknown function	YNR014W	CCTN	179	2.4	12.7	1.3
Ribosomal protein S27	RPS27B	CCTN	184	3.5	5.9	1.0
Abundant cell surface glycoprotein	SED1	CGAN	164	2.3	12.1	2.2
Alcohol dehydrogenase homolog	YMR318C	CGTN	101	5.0	44.4	0.8
S-adenosylmethionine synthetase	SAM1	CGTN	425	13.1	10.0	0.9
Plasma membrane Na^+-ATPase pump	ENA5	CTTN	107	12.4	6.6	1.1
UDP-glucose pyrophosphate	UGP1	CTTN	244	40.1	20.8	4.1
Mitochondrial chaperonin	HSP10	CTTN	293	no significant change	5.1	0.9

further contributed to the differences in gene expression profiles observed between the two studies. However, the discrepancy between the cDNA indexing and cDNA microarray data sets could not be solely attributed to these factors. Further evidence of this was provided by comparisons of the cDNA indexing data of this experiment to microarray data describing global gene expression changes in yeast following 20 minutes of saline treatment (TABLE 4.5). The correlation observed between cDNA indexing data and the microarray data detailing the 20-min timepoint was no better than that observed between the indexing data and the 10-min saline treatment microarray data.

4.3.15.3 Evaluation of 3'-end cDNA indexing of S. cerevisiae *for global gene expression profiling: progress summary*

The resolving capacity and sensitivity of detection of automated fluorescence-based DNA sequencing instrumentation for analysis of 3'-end cDNA indexing data provided a substantial improvement over analysis by agarose gel electrophoresis. Patterns of amplification products generated from independent parallel indexing ligations targeting the same complex population of *Fok*I-digested cDNA fragments were reproducible. The reproducibility of cDNA indexing profiles generated from distinct cDNA populations derived from parallel yeast cultures grown under identical conditions was also demonstrated. Distinct indexed gene expression profiles for yeast cultures grown in the presence of differing environmental stimuli were generated. Differences in the level of amplification of specific indexed 3'-terminal cDNA fragments were observed between the Salt and Glucose cDNA indexing data sets compared, suggesting the reporting of differences in the level of expression of specific mRNA transcripts between saline-treated and untreated yeast cultures. However, the difference in amplification levels between the Salt and Glucose data sets for specific 3'-terminal indexed cDNA fragments were found to be poorly correlated with data from published studies of saline shock response in *S. cerevisiae*.

The improved resolution and sizing of amplified fragments, and the heightened sensitivity of detection, afforded by analysis of cDNA indexing data sets with automated DNA sequencing instrumentation, revealed the presence of a surprisingly high number of unanticipated fragments among the Salt and Glucose indexed cDNA populations. Many of these unanticipated fragments were reproducibly amplified from the same cDNA population in parallel indexing reactions, and also from different cDNA sources. The presence of these unanticipated fragments raises concerns regarding their derivation, reduces the confidence with which specific indexed cDNA amplicons can be attributed to the representation of specific mRNA transcripts in the transcriptome of a particular yeast culture, and indicates the need for further refinement of the 3'-end cDNA indexing technique for effective application to global gene expression profiling in *S. cerevisiae*. Potential avenues of further research in this regard are outlined in the Discussion.

A number of anticipated indexed 3'-end cDNA fragments identified by the YeastORFdb database were not detected among the amplified transcripts in either the Salt or Glucose data sets (e.g. *DED81* in TABLE 4.5). Global gene expression data from numerous studies obtained by several different analytical approaches have indicated that, for most growth conditions, only 60%-70% of the yeast genome is expressed at any given time [228]. Even when this finding is taken into account, the absence of specific anticipated indexed fragments represents a significant challenge to the utility of cDNA indexing as an analytical method for global gene expression in *S. cerevisiae*. Possible explanations for the failure of cDNA indexing to detect these transcripts, and potential avenues of future research to improve their detection, are outlined in the Discussion.

The development of 3'-end cDNA indexing as a method for the analysis of global gene expression in *S. cerevisiae* has demonstrated promise, but has not yet been refined to the level required for efficient and fully informative application of the technique. Further development and optimization of cDNA indexing protocols will be necessary in order to capitalize on that promise. Several potential approaches for such further development will be discussed in the following section.

4.4 DISCUSSION

4.4.1 Development of 3'-end cDNA indexing protocols for global gene expression profiling in *S. cerevisiae*: summary

This investigation sought to develop modified protocols designed to facilitate global gene expression in *S. cerevisiae* by 3'-end cDNA indexing. Two alternate methods were employed for the synthesis of ds-cDNA populations from yeast total cellular mRNA in preparation for cDNA indexing analysis. Using transcript-specific primer pairs, the cDNA populations synthesized were demonstrated to be representative of the mRNA populations from which they were derived. *Fok*I digestion of these cDNA populations generated indexable 3'-terminal cDNA fragments predicted from ORF sequence data of the *S. cerevisiae* genome. The successful ligation of indexers complementary to cohesive end sequences of targeted *Fok*I-digested 3'-terminal cDNA fragments within a complex cDNA restriction digest was demonstrated by amplification of the indexed fragments using an indexing primer and transcript-specific primers. The selective capture of target fragments ligated to biotinylated indexers by streptavidin-coated paramagnetic beads was employed to reduce amplification reaction complexity. Stringent wash regiments eliminating nonspecific binding of nonbiotinylated cDNA to paramagnetic beads were established.

The efficiencies of several anchored poly(T) primer conformations for the amplification of indexed 3'-end cDNA fragments were evaluated using a series of artificial poly(A)-tailed indexable constructs. On the basis of this evaluation, anchored GC-rich poly(T) primers were selected to provide improved priming efficiency for cDNA population synthesis and for amplification. The artificial poly(A)-tailed indexable constructs were also used to evaluate the efficiency of 3'-terminal cDNA fragment amplification from biotinylated templates bound to streptavidin-coated paramagnetic beads, and to determine improved PCR cycling parameters for 3'-end cDNA indexing. The amount of particular transcript species needed for target amplification following the ligation of NoP indexers was evaluated using artificial indexed cDNA constructs. Ligation conditions providing high ligation fidelity with *Taq* DNA ligase were incorporated into the modified cDNA indexing protocol set.

Differential gene expression profiles for yeast cultures exposed to various environmental stimuli were identified using the modified cDNA indexing protocols. Expression of the *GAL1* transcript was observed in yeast grown in galactose-containing medium, while no *GAL1* expression was detected in yeast grown in glucose-containing medium. Increased expression of the *BOP3* transcript was observed in cDNA populations derived from pheromone-treated yeast cultures relative to cDNA populations derived from untreated cultures. These findings were well-correlated with published data obtained by established methods of gene expression analysis. A limited survey of gene expression changes in yeast responding to saline shock performed using a small number of indexers generated results compatible with published data obtained in studies of yeast salt shock response using cDNA microarrays.

Analysis of 3'-end cDNA indexing data by automated fluorescence-based DNA sequencing instrumentation was performed. The reproducibility of cDNA indexing profiles generated from independent parallel indexing ligations targeting individual cDNA populations and from distinct cDNA populations derived from parallel yeast cultures grown under identical conditions was demonstrated. Distinct indexed gene expression profiles were generated from cDNA populations derived from yeast cultures grown in the presence of differing environmental stimuli. Differences in the level of amplification of specific indexed 3'-terminal cDNA fragments were observed, indicating differences in the level of expression of specific mRNA transcripts between saline-treated and untreated yeast cultures. However, 3'-end cDNA indexing data sets were poorly correlated with data from published studies of saline shock response in *S. cerevisiae*. Unanticipated fragments were amplified, while a number of anticipated indexed 3'-end cDNA fragments were not detected in the cDNA indexing data sets analyzed. This indicates that the 3'-end cDNA indexing approach in its current form is not sufficiently refined for effective and informative application to global gene expression profiling in *S. cerevisiae*.

4.4.2 Future Research: addressing the challenges to 3'-end cDNA indexing identified in this investigation

The amplification of unanticipated fragments, and the failure to detect certain anticipated targets, reduces the confidence with which specific indexed cDNA amplicons can be attributed to the representation of specific mRNA transcripts in the transcriptome of yeast culture grown under particular environmental conditions, and indicates the need for further refinement of the 3'-end cDNA indexing technique for effective application to global gene expression profiling in *S. cerevisiae*. Possible causes of several of the unresolved challenges to 3'-end cDNA indexing are identified below. Potential avenues of future research through which solutions for these challenges might be discovered are outlined.

The amplification in cDNA indexing reactions of fragments not predicted from the YeastORFdb database might occur for several reasons. If PCR annealing conditions are not sufficiently stringent, GC-rich anchored poly(T) primers may anneal and misprime A/T-rich sequences of indexed non-3'-terminal cDNA fragments, permitting strand elongation toward the Bam-primer-binding sequence of the attached indexer. A single misprimed strand elongation event of this nature would generate a template efficiently primed in all subsequent cycles by the Bam and GC-rich anchored poly(T) primers. Further optimization of cDNA indexing PCR conditions to reduce mispriming while maintaining efficient amplification of legitimately-primed targets, or modification of GC-rich anchored poly(T) primer design, may inhibit the amplification of this class of unanticipated fragments.

Misligation of biotinylated indexers to noncomplementary cohesive end sequences borne by untargeted 3'-terminal cDNA fragments prior to streptavidin capture may also give rise to the amplification of unanticipated fragments in cDNA indexing reactions. The identification of more stringent ligation conditions requiring even higher ligation fidelity, without significantly sacrificing the efficiency of the current ligation protocol, would inhibit the amplification of unanticipated fragments by this mechanism. Another approach by which the same end might be achieved is the development of a cDNA indexing system based on the Type IIS restriction endonuclease *Tsp*RI (see **Section 1.2.1**). The unique cohesive end characteristics of fragments digested with this enzyme may effectively permit higher ligation fidelity using ligation conditions with efficiency similar to those currently employed.

Incomplete or unpredicted patterns of cDNA digestion may be the cause of unanticipated fragment amplification in specific rare cases. Digestion of a particular site of a specific cDNA transcript by the Type IIS endonuclease employed may be prevented in instances in which two recognition sites are closely associated in an appropriate orientation such that cleavage of the penultimate 3'-proximal site disrupts cleavage of the (anticipated) 3'-proximal site. This will result in the generation of an unanticipated 3'-end cDNA fragment of different length and bearing a different cohesive end sequence targeted by a different indexer. (This mechanism will necessarily contribute to the absence of predicted target fragments from the cDNA indexing data set.) The frequency of occurrence of this scenario, however, is likely to be rare. As a result, this mechanism is unlikely to be the major cause of the aberrant amplification patterns observed in this investigation.

One possible explanation for the failed detection of particular cDNA transcripts by cDNA indexing could be their inefficient synthesis from total cellular mRNA populations, as a result of bias due to secondary structure or poly(A) tail conformation. In addition to the infrequent instances attributable to this mechanism or to aberrant cDNA digestion, the absence of anticipated indexed fragments from a cDNA indexing data set may more commonly be due to competition within complex PCR mixtures. By this mechanism, accurately-digested and legitimately-ligated 3'-terminal cDNA fragments may exhibit a low relative amplification efficiency and be outcompeted by the amplification of other legitimate amplicons. This mechanism may be of particular importance in the failed amplification of rare or large transcripts from complex PCRs of many more common, smaller indexed targets. A possible solution for this, the exploitation of the "C_0t effect" in cDNA indexing PCRs, is discussed in **Section 4.4.3**.

4.4.3 Future Research: exploitation of the "C$_0$t effect" in future cDNA indexing applications

During the later cycles of PCR, a loss of amplification efficiency and a plateau of non-exponential amplification is observed which cannot be fully accounted for by the effects of competition, inhibition and reagent depletion. As amplification reactions progress, the concentration of PCR product strands rises sufficiently to allow the strands to increasingly anneal to each other during the time spent below the DNA melting temperature at each cycle. This reannealing of product strands may interfere with primer binding and thus with further amplification of these prominent PCR products. This "C$_0$t effect" [264-266] is especially prominent in cases in which more than one PCR product is being amplified. If significant levels of reannealing occur during the late cycles, abundant products will systematically be amplified less efficiently. Differences in abundance between initial starting templates in a single amplification reaction are diminished in the final products, permitting amplified concentrations of rarer transcripts to "catch up" to those of initially more abundant species. This effect may be exploited and enhanced to generate normalized product populations. Protocols for PCR which enhance normalization have been developed [264, 267, 268] which involve successive cycles of PCR with incrementally increasing reannealing temperatures and progressively lengthening reannealing and extension times. In this manner, abundant fragments would be expected to have a greater likelihood of reannealing, blocking primer binding sites on these fragments and favoring the amplification of (unblocked) rarer products of the complex PCR. This approach may be employed in the abundance normalization of single-tissue cDNA populations to raise sensitivity of 3'-end cDNA indexing to rare transcripts. Normalization of cDNA populations will reduce bias against the amplification of rare cDNA species. Further, the "C$_0$t effect" could be exploited in experiments comparing the expression states of two (or more) gene populations, such that cDNA species common to both populations anneal to themselves, permitting the biased amplification of cDNA species unique to one population or other. This "subtractive" approach would modify 3'-end cDNA indexing to permit differential applications with high sensitivity and bias towards detection of differentially-expressed genes.

Once a basic protocol for this technique was established, it could be adapted to a wide range of biologically-interesting applications, including studies of single cell types in various stages of development or stimulation by extracellular agonists; studies of the expression differences of closely- or distantly-related cell types; and studies of different pathological states within a single cell type. This last field of application is of particular interest, and has the potential to demonstrate the full range of versatility of 3'-end cDNA indexing, from qualitative first-pass diagnostics to more focused studies of progression of disease states in prostate cancer (for example) to identification of genes of particular interest. Linking this technique with other indexing-based technologies such as adaptor-tagged competitive PCR (ATAC-PCR) [215], and development of bioinformatics applications for these approaches, are of more long-term interest. An important long-term goal will be the automation of many steps of this integrated technology. Finally, an ultimate goal may be a response to the challenge presented by Eric Lander regarding the ability to monitor the expression of all genes simultaneously: "To decipher the logic of gene regulation we should aim to be able to monitor the expression level of all genes simultaneously, with a quantitative sensitivity level of less than one copy per cell, a qualitative sensitivity to distinguish all alternatively spliced forms [of mRNA], and the ability to assay single cells." [269]. With further research to address the challenges and capitalize on the promise identified in this current investigation, 3'-end cDNA indexing and its variants has the potential to become a powerful tool in the description and analysis of entire mRNA populations.

5 Chapter V - Conclusions and Future Work

5.1 CHAPTER SUMMARIES AND CONCLUSIONS

5.1.1 Development of DNA Indexing Strategies for Directed Mapping and Sequencing

The plasmid pUC19 was selected to form the basis of a model system for protocol development and troubleshooting of DNA indexing strategies. Optimal conditions for *Fok*I digestion of pUC19 DNA were established. The capability of DNA indexing to selectively amplify a specific target fragment from a DNA digest following ligation of cohesive-end-specific indexers to the target fragment was demonstrated. A PCR product was observed only when two indexers each complementary to one of the target's cohesive end sequences were present in the ligation. The pUC19 model system was indexed using the single-primer P/P indexing strategy in order to demonstrate the targeted amplification of each of the pUC19 *Fok*I fragments, and to observe the generation of non-targeted repeated-end fragments. Reaction conditions for the joining of indexers to target *Fok*I fragments by T4 DNA ligase were established to provide an appropriate balance between ligation fidelity and ligation efficiency. The ability of the P/NoP indexing strategy to eliminate the production of false positives due to repeated-end amplification was demonstrated.

Attempts to provide directionality in amplifications of indexed templates through the use of two indexer sets, each with a different core sequence and common primer, produced an artifact characterized as a nontypical form of primer-dimer (PD). This artifact was capable of out-competing the amplification of correctly-indexed target fragments.

The Bam and BamCC indexer sets were designed to target and amplify indexable fragment classes in a manner that imported end-specific priming sites but that avoided primer-dimer production. The compound-primer P/NoP indexing strategy incorporated the P/NoP approach preventing the amplification of repeated-end fragments, provided the means for directional cycle sequencing of amplified indexed fragments, and simultaneously eliminated PD artifact amplification from indexing PCR reactions. The effective use of the Bam/BamCC indexing sets to perform compound-primer P/NoP indexing was demonstrated by the efficient indexing and amplification of each of the pUC19 target fragments. No misligation products were amplified following the development and application of indexing protocols providing high ligation fidelity without sacrificing ligation efficiency.

Each of the five pUC19 *Fok*I fragments was indexed and amplified twice, with opposite directionality, using the Bam/BamCC compound-primer P/NoP approach. Direct cycle sequencing of the amplified indexed pUC19 fragments was performed. Alignment of indexing-based directionally-sequenced pUC19 templates to an indexing-based restriction map of pUC19 constructed by jigsaw assembly.

The pUC19 model system was successfully exploited in the development of indexing strategies and appropriate indexing protocols. Model system characteristics presenting challenges to indexing approaches mimicking those presented in more complex systems were investigated, and indexing strategies developed to meet those challenges. The compound-primer strategy was successfully applied to the pUC19 model system, in a manner that permitted the complete characterization and sequencing of the system.

On the basis of these results and the principles fundamental to DNA indexing approaches, a proposal for a future application of DNA indexing is presented (see **Section 5.2**). A model for the use of multiplex indexing in the identification of Type IIS cohesive end sequences is described. The use of indexing sequence tagged sites, generated from pairs of indexed S/F fragments centered on *Sfi*I restriction sites, for physical mapping of complex genomes is outlined. Finally, an entirely indexing-based method for efficient directed mapping and sequencing of prokaryotic genomes is proposed.

5.1.2 Bacterial Strain and Species Differentiation by Indexed Genomic Profiling (IGP)

An indexing-based approach to microbial molecular subtyping was developed and demonstrated by adapting existing indexing protocols for the complexity of microbial genome analysis. The use of pools of NoP indexer sequences in ligations of bacterial genomic digests provided an efficient method of indexing for profiling applications. Ligation conditions for bacterial profiling were established through the evaluation of indexer concentration, DNA concentration, and ligase concentration. *Taq* DNA ligase and *E. coli* DNA ligase were evaluated as IGP indexing ligases against T4 DNA ligase. Further development of reaction conditions for IGP ligations employing T4 DNA ligase established reaction parameters including incubation temperature, incubation time and ligase concentration. Software was developed to facilitate manipulation of bacterial genome sequence data for DNA indexing analysis. Initial application of the modified protocols to the molecular fingerprinting and differentiation of several *E. coli* strains was accompanied by predictive modelling based on the published genomic DNA sequence of *E. coli* strain MG1655. The ability of IGP studies to differentiate between related bacterial strains using small numbers of indexer combinations was demonstrated. Ligations generating legitimate indexed amplicons from predicted *Fok*I target fragments were amplified without substantial levels of background amplification products. Molecular subtyping of three common laboratory strains of *E. coli* was performed by indexed genomic profiling using complete sets of NoP indexer mixes. Indexed genomic profiles were generated from clinical isolates and reference strains of several *Staphylococcus* species. Prior determination of genomic sequence data was not necessary for the generation of specific *Staphylococcus* species profiles, presenting information regarding fragment size and cohesive end sequence for hundreds of profile datapoints across the species profiled. Indexed genomic profiling was found to provide excellent discriminatory power in the form of an information-dense molecular fingerprint derived by objective sampling of microbial genetic structure. Potential application of IGP to epidemiological studies of pathogenic

Staphylococcus species and strains, and to studies of microbial community diversity, were discussed.

5.1.3 Global Gene Expression Profiling of *Saccharomyces cerevisiae* by 3'-end cDNA Indexing

Modified 3'-end cDNA indexing protocols were developed to facilitate global gene expression profiling in *S. cerevisiae*. *Fok*I digestion of cDNA populations generated indexable 3'-terminal cDNA fragments predicted from ORF sequence data of the *S. cerevisiae* genome. Indexers were ligated to the complementary cohesive end sequences of targeted *Fok*I-digested 3'-terminal cDNA fragments within a complex cDNA restriction digest. The selective capture of target fragments ligated to biotinylated indexers by streptavidin-coated paramagnetic beads was employed to reduce amplification reaction complexity, utilizing stringent wash regimens to eliminate the nonspecific binding of nonbiotinylated cDNA. Artificial poly(A)-tailed indexable constructs were employed in the evaluation of anchored poly(T) primer conformations, determination of the efficiency of 3'-terminal cDNA fragment amplification from templates bound to paramagnetic beads, identification of the amount of particular transcript species needed for amplification of indexed targets, and the establishment of improved PCR cycling parameters. Modified cDNA indexing protocols were employed to generate differential gene expression profiles for yeast cultures exposed to various environmental stimuli. Differential expression of the *GAL1* and BOP3 transcripts were observed, and the results correlated to data obtained by other analytical approaches. A limited survey of gene expression changes in yeast responding to saline shock performed using a small number of indexers generated results compatible with published data obtained in studies of yeast salt shock response using cDNA microarrays.

Analysis of 3'-end cDNA indexing data by automated fluorescence-based DNA sequencing instrumentation revealed the reproducibility of cDNA indexing profiles generated from independent parallel indexing ligations targeting individual cDNA populations and from distinct cDNA populations derived from parallel yeast cultures grown under identical conditions. Distinct indexed gene expression profiles were generated from cDNA populations derived from yeast cultures grown in the presence of differing environmental stimuli. Differences in the level of amplification of specific indexed 3'-terminal cDNA fragments were observed, indicating differences in the level of expression of specific mRNA transcripts between saline-treated and untreated yeast cultures. However, 3'-end cDNA indexing data sets were poorly correlated with data from published studies of saline shock response in *S. cerevisiae*. Unanticipated fragments were amplified, and certain anticipated indexed 3'-end cDNA fragments were not detected in 3'-end cDNA indexing data sets, indicating that refinement of the 3'-end cDNA indexing technique is necessary for effective application to global gene expression profiling in *S. cerevisiae*. Avenues of future research were identified that may provide solutions to unresolved challenges to cDNA indexing approaches.

5.2 Future Developments: Indexing-Directed Bacterial Genomics (IDBG)

5.2.1 Identification of Type IIS cohesive end sequences by multiplex indexing

Ordered map construction from indexing template sequences by jigsaw assembly, and indeed the targeted amplification of a particular restriction fragment of interest, requires knowledge of the cohesive end sequences of the indexing target fragments involved. In cases where the sequence of the DNA being analyzed by indexing is not known, a means of identifying the cohesive end sequences of fragments generated by Type IIS restriction endonucleolytic digestion is necessary. Direct fluorescent sequence analysis of Type IIS fragment ends, employed by Brenner and Livak to facilitate contig mapping [270] is limited in application to the analysis of 5'-overhanging cohesive ends present in simple DNA digests. An alternative, general method for decoding the cohesive end sequences of fragments in DNA digests of significant complexity is multiplex indexing [271]. A description of this approach follows.

Multiplex indexing of the 4-nt cohesive end sequences of *Fok*I restriction fragments incorporates the use of eight ordered mixtures of indexers, or *multiplex* *mixes* (TABLE 5.1). Each multiplex mix contains exactly 128 nonphosphorylated BamCC indexers, representing half of the possible 4-nt cohesive end sequences. These mixes are composed in such a way that each particular four-base sequence appears in a unique combination of mixes. As there are exactly 256 possible patterns of presence (+) or absence (-) of a single indexer distributed uniquely across 8 mixes, the pattern of presence or absence of a particular indexer is equivalent to a binary code for each cohesive end sequence. The same is true for the presence or absence in a PCR reaction of an amplified product corresponding to a fragment indexed at one end by that particular indexer.

In a multiplex indexing experiment, a single phosphorylated Bam indexer is added to each of ten ligations to permit amplification of fragments bearing sequence complementary to that indexer on one of their cohesive ends. (This end sequence will be referred to as the "known" end, as only fragments bearing that sequence will be amplified in that reaction.) As each of the 128 indexers present in the multiplex mix added to the ligation is nonphosphorylated, only fragments with the P-indexer ligated to one end and a NoP indexer present in the multiplex mix ligated to the other can be amplified. Amplification of an indexed product occurs only in multiplex mix ligations that contain the NoP indexer complementary to the "unknown" cohesive end sequence of the fragment. The composition of the multiplex mixes permits that gel-analyzed results of indexed PCRs can be readily interpreted to provide the cohesive end sequence of the fragment's "unknown" end. Each pair of the 8 multiplex mixes codes for one base position of the cohesive end sequence. The pattern of product vs. no-product for a fragment of a particular size observed in 8 multiplex indexed amplifications electrophoresed in 8 lanes of an agarose gel will correspond to the binary code for the NoP indexer required for that fragment's specific targeted amplification. The binary coding pattern for each of the four bases (A, C, G or T) present in the multiplex mixes is shown in TABLE 5.2. The code is readily interpretable as representing the indexer sequence required to complement the unknown cohesive end sequence of the target fragment, which provides direct information about the fragment's end sequence. For example, a DNA fragment bearing

TABLE 5.1: Multiplex mix composition for cohesive end identification.
(modified from *Carlisle et al.* [271])

Mix#	NoP Indexer Cohesive Ends
1	(G,T)NNN==========
2	(C,T)NNN==========
3	N(G,T)NN==========
4	N(C,T)NN==========
5	NN(G,T)N==========
6	NN(C,T)N==========
7	NNN(G,T)==========
8	NNN(C,T)==========

N represents a mixture of all 4 bases, and ========== represents the double-stranded indexer core sequence.

TABLE 5.2: Decoding of binary multiplex signals to identify cohesive ends.
(modified from *Carlisle et al.* [271])

	A	C	G	T
Multiplex Signal	- -	- +	+ -	+ +

The "signals", or patterns of amplified indexed fragments across the lanes of a multiplex analysis gel, are read in pairs: (5')-mixes (1, 2); (3, 4); (5, 6); (7, 8)-(3'). "+" indicates the presence of a specific PCR product, while "-" denotes the absence of a PCR product of that size. The decoded sequence corresponds to the indexer cohesive end sequence complementary to the target fragment's unknown cohesive end sequence.

the sequence GCTA on its unknown end would be amplified only from multiplex ligations containing the complementary indexer sequence TAGC. This indexer sequence is present in multiplex mixes 1, 2, 5, and 8. Amplification of indexed product of the appropriate size would thus be visualized as bands in lanes 1, 2, 5 and 8 of an 8-lane gel. In other words, the "signals" from the 8 multiplex ligations would be + + (T)/ - - (A)/ + - (G)/ - + (C).

Two indexer sequences require special consideration in order to ensure correct interpretation of multiplex indexing data. Fragments bearing the sequence TTTT on their unknown end are amplifiable following ligation of the NoP BamCC indexer AAAA to that end. Due to the composition of the multiplex mixes to provide a readable binary code, this indexer is not represented in any one of the eight indexer mixtures. As a result, the complete absence of a PCR product in all lanes of a multiplex experiment may reflect the actual absence of any indexing target, or the potential presence of a target (of undefined size) bearing a TTTT end sequence. To avoid this ambiguity, a ninth ligation containing the P-indexer and the NoP indexer OH-AAAAxBamCC alone is performed. The presence of a fragments with a TTTT end sequence that is accurately coded by the multiplex reactions (i.e. is unamplified in the first eight reactions) will be amplified in the ninth, providing positive confirmation of the unknown end sequence and also the size of the target fragment. A similar situation arises in the instance of fragments bearing the end sequence AAAA at their unknown end. The amplification of these products in all eight lanes of the multiplex reactions, due to the presence of the complementary indexer OH-TTTTxBamCC in every multiplex indexer mixture, is indistinguishable from the spurious amplification of repeated-end fragments bearing the sequence targeted by the P-indexer in every reaction. The addition of a tenth multiplex reaction containing only the P-indexer as the sole indexing sequence differentiates these two possibilities: if no PCR product identical in size to the putative AAAA-fragment amplicon is obtained in lane 10, then the 8-lane multiplex data is correct. If a fragment identical in size to that amplified in all 8 lanes is amplified in lane 10, then the multiplex data for that fragment size may be ignored as a confirmed false positive.

In some envisioned applications of multiplex indexing (see below), complex digests are aliquoted into 256 different ligation reactions, each containing a different biotinylated P-indexer. The addition of streptavidin-coated paramagnetic beads to each of these ligations allows the biotinylated indexer sequences to bind to the beads. The beads from each ligation extraction are washed, removing unbound DNA fragments and retaining fragments ligated at one end to the specific P-indexer sequence represented in that tube. In multiplex indexing experiments using fragments prepared in this manner, the description of the P-indexer-targeted end of the fragment as the "known" end is particularly appropriate: the fragment would not be retained in the tube corresponding to that particular indexer sequence unless it was ligated to the P-indexer. In these experiments, the likelihood of ambiguous multiplex coding results for unknown ends due to false positives arising from repeated-end fragments targeted by the P-indexer in that reaction is further reduced due to the effects of the beads on the availability of free P-indexer for both ends of the same molecule.

5.2.2 Indexing sequence-tagged site (iSTS) mapping

Outside of an indexing context, the term *sequence tagged site* (STS) refers to a small (~200-300 bp) DNA region targeted by a pair of primers permitting specific amplification of that unique sequence from a known chromosomal location identifying the position of a gene. Targeting of random and functionally-neutral DNA sequences such as microsatellites by primer pairs allow STSs to act as physical markers for genomic mapping and cloning. STSs present on a particular clone can be identified and ordered along STS-based physical maps of chromosomes or genomes, linked to genes on a genetic map, as well as being detected on a radiation hybrid map. This facilitates the alignment of the various map types and assists in the ordering of YAC, BAC or P1 clone contigs which may be employed in a genomic mapping or sequencing effort.

The use of pairs of indexed fragments centered on *Sfi*I restriction sites for *indexing sequence-tagged site* (iSTS) mapping has been described by Unrau and Deugau [22, 272]. The IP restriction endonuclease *Sfi*I recognizes the interrupted palindromic sequence 5'-GGCCN$^{\mathrm{v}}$NNNNGGCC-3' and cleaves substrate DNA in a manner that generates 3-nt 3'-overhanging informative cohesive ends. As there are 64 different trinucleotide sequences, use of this enzyme for indexing purposes generates 64 cohesive-end classes and 2080 fragment classes. *Sfi*I is classified as a "rare cutter", as the eight-base specificity of its recognition sequence is expected to occur only once every 65 536 (on average) bp in random-sequence DNA with 50% G/C content. Using Type IIS endonucleases which cleave frequently in a particular DNA sequence (such as *Fok*I, with a recognition sequence occurring every 512 bp on average) in combination with rare cutters like *Sfi*I generates indexable fragments which define iSTS centered on the rare cutting sites.

For the purposes of indexing-based physical mapping, iSTSs have the form F_L-S-F_R, where F_L and F_R are Type IIS (e.g. *Fok*I) sites and S is the rare cutter (e.g. *Sfi*I) site. In other words, each iSTS is composed of a *Fok*I fragment, indexable on its two 4-nt 5'-cohesive ends, bearing an *Sfi*I restriction site between its left and right *Fok*I cut sites. Following cleavage of the *Fok*I fragment with *Sfi*I, two *Sfi*I/*Fok*I (S/F) fragments are generated which are indexable on each of their 3-nt 3'-cohesive ends. As a result, each iSTS may be characterized by total (*Fok*I fragment) length, left-side S/F and right-side S/F fragment lengths, and a total of 11 bases of informative cohesive end sequence. When compared to the low frequency of *Sfi*I sites (roughly 80 sites in a typical 5 Mb bacterial genome, or about 40 000 in the human genome), the large number of possible cohesive end sequences (256×64 or 16 384) defining a single S/F fragment suggests that any one pair of indexers targeting the *Sfi*I and *Fok*I cohesive ends of an iSTS S/F fragment will index only one or two fragments in even the human genome. As one pair of indexers targets only the right or the left side of the iSTS site (i.e. only the right-side S/F fragment or the left-side S/F fragment), two pairs of indexers are used to define the two S/F fragments of the complete iSTS. Therefore for indexing-based physical maps constructed with *Fok*I and *Sfi*I, there are $2.1\text{x}10^6$ possible iSTS classes derived by cohesive end sequence alone, independent of fragment length information. (Further iSTS information is derived from the observed lengths of amplified indexed fragments. S/F fragment lengths may be used to differentiate between two iSTSs bearing identical sequence information, in the rare event that such discrimination is necessary.) Even greater specificity is expected in iSTS mapping systems employing the 5-nt-end-generating *Hga*I as the frequent cutter, offering 256 x 10^9 unique iSTS cohesive end sequence classes. In any prokaryotic or mammalian genome mapped using indexing-based techniques, therefore, each iSTS is likely to be unique and definitive for physical map construction.

5.2.3 Indexing-Directed Bacterial Genomics (IDBG): an efficient non-cloning method for the directed mapping and sequencing of prokaryotic genomes

I describe a potential strategy for complete bacterial genomic physical mapping and directed sequencing by DNA indexing (shown schematically in FIGURE 5.1). The method outlined is completely non-cloning, and is therefore not subject to the problems intrinsic to cloning including selective fragment loss, chimera formation, mutation and sequence rearrangement. This method is expected to be cost-efficient and is amenable to automation. The example of a hypothetical moderately large (5 Mb) bacterial genome is used to illustrate the manner in which this method may be applied.

5.2.3.1 Sfi-series indexer sets for genomic mapping and sequencing

Three related sets of indexers are required for the IDBG approach to bacterial genomic mapping and sequencing (FIGURE 5.2). These sets are designed such that fragments targeted by indexers from any combination of the three indexer sets may be amplified using a single common primer. The SfiCA 3'3b biotinylated P-indexer set (FIGURE 5.2A) is designed to target the 3-nt 3' cohesive end sequences of *Sfi*I fragments. Indexers targeting 3' overhangs bear their specific indexing sequence on the same strand as their primer sequence. Each of the 64 indexing/priming strands employed in this set is 5'-biotinylated to enable indexed-fragment capture by streptavidin-coated paramagnetic beads. The 5'-phosphorylated strand of the SfiCA P-indexers is composed of the common Sfi primer-binding sequence shared by all three

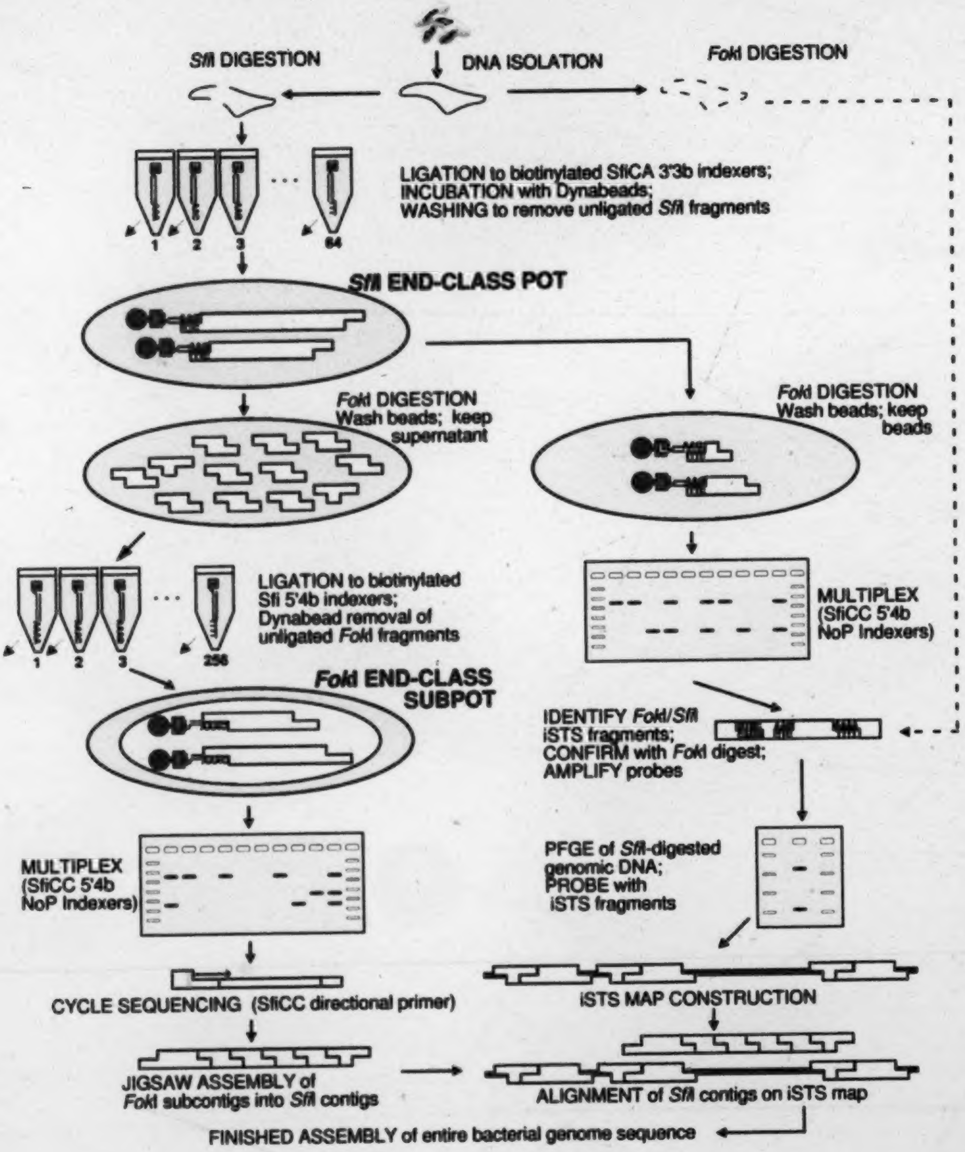

SfiI DIGESTION
DNA ISOLATION
FokI DIGESTION
LIGATION to biotinylated SfiCA 3'3b indexers;
INCUBATION with Dynabeads;
WASHING to remove unligated SfiI fragments
1
2
3
...
64
SfiI END-CLASS POT
FokI DIGESTION
Wash beads; keep supernatant
FokI DIGESTION
Wash beads; keep beads
LIGATION to biotinylated Sfi 5'4b indexers;
Dynabead removal of unligated FokI fragments
1
2
3
...
256
MULTIPLEX (SfiCC 5'4b NoP Indexers)
FokI END-CLASS SUBPOT
IDENTIFY FokI/SfiI iSTS fragments;
CONFIRM with FokI digest;
AMPLIFY probes
MULTIPLEX (SfiCC 5'4b NoP Indexers)
PFGE of SfiI-digested genomic DNA;
PROBE with iSTS fragments
CYCLE SEQUENCING (SfiCC directional primer)
iSTS MAP CONSTRUCTION
JIGSAW ASSEMBLY of FokI subcontigs into SfiI contigs
ALIGNMENT of SfiI contigs on iSTS map
FINISHED ASSEMBLY of entire bacterial genome sequence

FIGURE 5.1: Schematic flow diagram of Indexing-Directed Bacterial Genomics
Only *FokI*-based subcontig mapping is shown. *Sfa*NI-based subcontig mapping is to proceed in a similar manner. See text for other details.

A) SfiCA 3'3b Biotinylated P-Indexer Set

Indexing strand sequence:

5'-*Biot*-GGTACAGAGTCTACGAAGGCCANNN-3'	indexing oligo (25 nt)

Features:

GGTACAGAGTCTACGAAGGC	primer (PCR) (20 nt)
GGTACAGAGTCTACGAAGGC***CC***	directional primer can't primek mismatch (22 nt)
GGCCaNNN	*Sfi* site (one half)

Core strand sequence:

5'-***P***-TGGCCTTCGTAGACTCTGTACC-3'	phosphorylated (22 nt)

B) Sfi 5'4b P-Indexer Set

Indexing strand sequence:

5'-***P***-NNNNGCCTTCGTAGACTCTGTACC-3'	indexing oligo (24 nt)

Primer strand sequence:

5'-GGTACAGAGTCTACGAAGGC-3'	nonphosphorylated (20 nt)

C) SfiCC 5'4b NoP Compound Indexer Set

Indexing strand sequence:

5'-***P***-NNNN***GG***GCCTTCGTAGACTCTGTACC-3'	indexing oligo (26 nt)

(Directional) primer strand sequence:

5'-GGTACAGAGTCTACGAAGGC***CC***-3'	nonphosphorylated (22 nt)

FIGURE 5.2: Features of the three proposed Sfi Indexer sets.

related sets of indexers, in addition to the CA dinucleotide to permit them to be annealed to the SfiCA indexing strands leaving only a 3-nt single-stranded overhang. The presence on the annealed double-stranded SfiCA indexers of one-half of the *Sfi*I recognition sequence permits re-cleavage of SfiCA indexers from target fragments following ligation. (The other half of the palindromic *Sfi*I site is necessarily present on the targeted *Sfi*I restriction fragment, as it permitted digestion at that location in the first place.)

The second set of indexers employed in the IDBG approach is the Sfi 4b5' biotinylated P-indexers (FIGURE 5.2B). These indexers target the 4-nt 5' cohesive end sequences generated by *Fok*I or *Sfa*NI restriction endonucleases. The indexing strands of these indexers are 5'-phosphorylated, bear one of the 256 4-nt sequences on their 5' end, and contain the 20 nt common Sfi primer-binding sequence. The primer strand is the Sfi common primer. A biotinylated form of the primer strand is employed in the construction of Sfi biotinylated P-indexers, while the unmodified primer strand is used as a primer for PCR amplification of indexed fragments.

The third indexer set, the SfiCC 4b5' NoP compound indexers (FIGURE 5.2C), is the indexer format found in the multiplex mixes employed in this proposed indexing approach. The SfiCC NoP compound indexers and the Sfi P-indexers act in an analogous manner to the BamCC NoP compound indexers and Bam P-indexers. Both SfiCC and Sfi indexers target the same cohesive end format, permit the amplification of indexed targets through PCR with their common primer, enable P/NoP indexing, and provide directionality in cycle sequencing via the extra CC dinucleotide on the corresponding directional sequencing primer. The SfiCC directional sequencing primer does not accurately base pair to SfiCA indexers, and therefore allows directional sequencing of indexed templates generated through ligation with either of the two related P-indexer sets SfiCA or Sfi. In contrast to SfiCA indexers, neither SfiCC nor Sfi indexers are re-cleaved from indexed target fragments by digestion with *Sfi*I endonuclease. Though half of the palindromic *Sfi*I restriction site is present on SfiCC indexers (it is absent in the shorter Sfi P-indexers), the restriction fragments to which SfiCC indexers are ligated are not generated by *Sfi*I digestion, and thus do not carry the other half of the palindromic sequence required for recognition and digestion.

5.2.3.2 Classification and isolation of SfiI *fragments based on cohesive end sequence*

Bacterial genomic DNA is purified from cultures of the organism of interest. A small aliquot (10 µg) of genomic DNA is digested with *Fok*I and held in reserve for later use in the indexing and amplification of STS fragments. The bulk of the genomic DNA preparation is digested with *Sfi*I endonuclease. In a typical 5 Mb bacterial genome, *Sfi*I cleavage is expected to generate about 80 fragments averaging 65.5 kb in length. As each fragment carries two cohesive end sequences, 160 different cohesive ends would be present in such a digest.

Following *Sfi*I digestion of the bacterial genome, the digested DNA is apportioned into 64 aliquots. One aliquot of digested DNA is added to each of 64 ligations, each reaction containing one of the 64 SfiCA 3b3' biotinylated P-indexers and a DNA ligase. (T4 DNA ligase may be employed, or if desired ligation may be performed with *Taq* DNA ligase as described by Shaw-Smith *et al.* [223] and discussed in **Chapter IV**.) If a 5-Mb bacterial genome contains approximately 80 *Sfi*I restriction fragments, representing 160 different cohesive ends, then each of the 64 SfiCA indexer sequences might be expected to target between 2 and 5 different fragment ends in any one ligation. In other words, each reaction will contain between 2 and 5 fragments ligated to the specific biotinylated SfiCA indexer present. As each restriction fragment has two indexable ends, each fragment will be ligated to indexers in two separate reactions. Addition of streptavidin-coated paramagnetic beads to each ligation permits extraction of the biotinylated indexer, and of the *Sfi*I fragments ligated to it. Other restriction fragments are removed, leaving 2 to 5 *Sfi*I fragments, ligated to the indexer sequence complementary to one of their cohesive ends, in any particular well of the 64-well experiment. Each of the 64 reactions, containing 2 to 5 *Sfi*I fragments ligated at one end to complementary SfiCA indexers, will be referred to as an *SfiI end-class pot*.

For the fragments present in each *Sfi*I end-class pot, several steps need to be performed. First, each of the *Sfi*I fragments in the end-class pot must be ordered along the bacterial chromosome, generating a physical map of the genome based on iSTSs. The *Sfi*I fragments in each pot must be cleaved by *Fok*I, the *Fok*I fragments multiplex-indexed and amplified to facilitate mapping along the *Sfi*I "contig", and the indexed *Fok*I amplicons used as templates for cycle sequencing. This step may then be repeated in the instances of certain *Sfi*I end-class pots using *Sfa*NI in place of *Fok*I as the secondary indexing restriction enzyme. The *Fok*I (and *Sfa*NI) template sequences must be assembled along the *Sfi*I fragment structure into *Sfi*I-fragment contigs. The final step required for the complete mapping and sequencing of a bacterial genome by DNA indexing is the assembly of sequenced *Sfi*I-fragment contigs along the scaffold of the iSTS physical map.

5.2.3.3 Assembly of the iSTS physical map

Assembly of a physical genomic map proceeds by the identification of iSTSs representing junctions between adjacent *Sfi*I fragments along the chromosome. In preparation for iSTS identification, a small aliquot of the paramagnetic bead suspension (1/25 of the total suspension volume) is removed from each of the 64 tubes of the *Sfi*I end-class pot set. (The remaining volume in the 64 bead-suspension tubes is stored for later use in subcontig assembly and sequencing.) The *Sfi*I fragments attached to paramagnetic beads in each (1/25 volume) *Sfi*I end-class pot are digested with *Fok*I endonuclease. After the beads are washed to remove unbiotinylated *Fok*I restriction fragments, the only fragments remaining in a particular *Sfi*I end-class pot are the short *Sfi*I/*Fok*I (or S/F) fragments produced by *Fok*I cleavage of each *Sfi*I fragment at the *Fok*I site most proximal to the biotinylated indexed *Sfi*I cohesive ends. Each *Sfi*I fragment present in a particular *Sfi*I end-class pot is represented by a single S/F fragment of unique length and *Fok*I cohesive end sequence. (At this stage of the procedure, however, the 4-nt *Fok*I cohesive end sequence and length of each S/F fragment are unknown, as is the actual number of S/F fragments in each *Sfi*I end-class pot).

The number of S/F fragments present in each pot (reflecting the number of *Sfi*I restriction fragments with a particular 3-nt *Sfi*I cohesive end sequence), the length of those fragments and each of their *Fok*I cohesive end sequences are determined by multiplex indexing. Ten equal aliquots of each *Sfi*I end-class pot bead suspension are added to ten ligation reactions composing a single multiplex indexing reaction set (one set of ten multiplex indexing reactions for each *Sfi*I end-class pot, for a total of 640 multiplex indexing ligations). In each multiplex set, reactions 1 to 8 contain multiplex indexer mixes composed of SfiCC 4b5' NoP compound indexers present in the cohesive end sequence combinations described in **Section 5.2.1**. Multiplex reaction 9 contains the single indexer OH-AAAAxSfiCC, acting as a positive control for the (negative) AAAA multiplex signal, as previously described. For the IDBG multiplex indexing of S/F fragments, multiplex reaction 10 contains a mixture of all 256 SfiCC 4b5' NoP indexers, not the single indexer OH-TTTTxSfiCC as would be expected for the multiplex indexing of *Fok*I restriction fragments. The 256-indexer mix is used to allow amplification of all S/F fragments in a particular *Sfi*I end class pot in a single reaction for use as size standards and as a positive control. [The use of the single TTTT indexer in *Fok*I fragment multiplex indexing is to differentiate between informative multiplexing of fragments with AAAA cohesive end sequences (a positive signal in all of multiplex reactions 1 to 8) and repeated-end fragments bearing the complementary sequence to the P-indexer used in the multiplex reactions. S/F fragments have only one 4-nt 5'-cohesive end, and the biotinylated P-indexer used to capture the fragments targets only *Sfi*I cohesive ends. The use of NoP indexers for *Fok*I cohesive end targeting ensures that no repeated-end fragment amplification is possible in S/F multiplex indexing.]

Following amplification of the S/F fragments in the multiplex reactions using the Sfi common primer, the multiplex reaction products are analyzed by gel electrophoresis. If, as anticipated, there are between 2 to 5 *Sfi*I fragments in the genomic *Sfi*I digest which bear a particular sequence on one of their 3'-cohesive ends, digestion with *Fok*I generates an equal number of S/F fragments, each bearing a particular *Fok*I cohesive end sequence and being of a particular defined length. The

*Fok*I cohesive end sequences of each S/F fragment is determined by the signal pattern of amplified product across the multiplex indexing reactions, and the length is determined by electrophoretic separation on the gel. Even in rare instances in which two S/F fragments with the same *Sfi*I-end sequence also share the same *Fok*I-end sequence, it is likely that the lengths of the two S/F fragments will be different and therefore resolvable by electrophoretic analysis.

For example, if there are 3 *Sfi*I fragments in the 5 Mb genome that bear the sequence GTT on one cohesive end, they will be ligated to the biotinylated indexer P-AACxSfiCA in *Sfi*I end-class pot 2 (of 64) and bound to the paramagnetic beads. Digestion of captured *Sfi*I fragments by *Fok*I generates 3 S/F fragments in pot 2, one for each *Sfi*I fragment: a 1020-bp fragment with *Fok*I cohesive end CACC; a 300-bp fragment with *Fok*I cohesive end TTTT; and a 250-bp fragment with *Fok*I cohesive end TCCG. Multiplex indexing of pot 2 will generate a GGTG signal (+-/+-/++/+-) at 1064 bp on the gel, an AAAA signal (–/--/--/--/+) at 344 bp, and a CGGA signal (-+/+-/+-/--) at 294 bp. (The shift in size is due to the addition of indexers to each end of the fragments.)

Each S/F fragment represents one-half of an iSTS centered on an *Sfi*I cut site. For each S/F fragment with a particular *Sfi*I end sequence, there is another S/F fragment bearing a complementary *Sfi*I end sequence. All 3-nt sequences are non-palindromic. Consequently, the number of S/F fragments in a particular *Sfi*I end-class pot is equal to the number of S/F fragments in the *Sfi*I end-class pot defined by the complementary *Sfi*I end sequence. For instance, if there are three S/F fragments in the AAC-defined *Sfi*I end-class pot 2, then there should be three S/F fragments in the GTT-defined *Sfi*I end-class pot 48. (If there are not, a problem requiring specific intervention is indicated.) Multiplex indexing of pot 2 provides the lengths and *Fok*I end sequences of the three AAC-defined S/F fragments; multiplex indexing of pot 48 provides the lengths and *Fok*I end sequences of the three GTT-defined S/F fragments. The correct pairing of S/F fragments from each pot is necessary to identify and amplify *Fok*I fragments which may be used as iSTSs.

In order to correctly pair S/F fragments from complementary *Sfi*I end class pots, ligations of *Fok*I-digested genomic DNA are prepared in which each identified *Fok*I sequence from pot 2 is matched with each of the identified *Fok*I sequences from pot 48 and targeted by appropriate indexers. (If each end class pot contains 3 S/F fragments, then nine ligations targeting each of the putative *Fok*I fragments are required.) A 5 Mb genome is expected to contain about 10 000 *Fok*I fragments. As less than one-third of the 32 896 *Fok*I fragment classes will be represented at all in such a digest, the probability that any two 4b5' indexers will amplify more than one fragment is small. In instances in which two or more PCR products are present in a particular indexing reaction targeting a putative *Sfi*I-fragment-linking *Fok*I fragment, the correct (iSTS) fragment may be identified on the basis of size (which should match the expected length of a complete fragment assembled from the lengths of the two underlying S/F fragments). Further confirmation of iSTS *Fok*I fragment identity is obtained by digesting the amplified *Fok*I fragment with *Sfi*I, generating products of identical size to the each of the S/F fragments used to generate the iSTS *Fok*I end sequences. Final validation is provided by ligation of the appropriate 3b3' indexers to the digestion products followed by amplification, thus demonstrating that the putative iSTS *Fok*I fragment bears an *Sfi*I restriction site generating the correct sequence in the correct location along its length.

Identification and amplification of the complete set of iSTS *Fok*I fragments crossing the junctions of adjacent *Sfi*I fragments permits those iSTS fragments to be used as a radiolabeled or fluorescently-labeled probes in the ordering of *Sfi*I fragments for physical map assembly. Pulsed-field gel electrophoresis (PFGE) of *Sfi*I-digested genomic DNA fragments is performed to separate *Sfi*I fragments on the basis of size. The gel is probed with each labeled iSTS *Fok*I fragment in turn, identifying which *Sfi*I fragments are linked by that iSTS fragment and are therefore adjacent to one another along the chromosome. This information, in addition to the *Sfi*I fragment lengths, is used to construct an iSTS physical map of the prokaryotic genome which acts as a "scaffold" along which sequenced *Sfi*I fragment contigs may be aligned.

5.2.3.4 Fok*I-based subcontig mapping and sequencing of* SfiI *fragments*

The remaining volume in each of the original 64 bead-suspension tubes is divided into two equal aliquots. One set is used for *Fok*I-based subcontig mapping and sequencing of each *Sfi*I fragment, and the other is used for *Sfa*NI-based subcontig mapping and sequencing. For *Fok*I-based mapping and sequencing template generation, the bead-captured *Sfi*I fragments in each *Sfi*I end-class pot is digested with *Fok*I endonuclease. The bead suspensions are washed once, and the beads (carrying the biotinylated S/F fragments) are removed, leaving in each *Sfi*I end-class pot a population of *Fok*I restriction fragments derived from the 2 to 5 *Sfi*I fragments initially present in the pot. Each of the 64 *Fok*I fragment populations are subdivided into 256 aliquots. Each aliquot is added to a ligation reaction containing one of the 256 Sfi biotinylated P-indexers. Ligation of the biotinylated 4b5' indexer in a particular *Fok*I *end-class subpot* to fragments bearing the cohesive end sequence targeted by that indexer permits the isolation of targeted fragments in each subpot. Streptavidin-coated paramagnetic beads are added to the completed ligation reactions, incubated, and unbound DNA removed. Each subpot thus contains all *Fok*I fragments (bearing a particular *Fok*I end sequence) derived from the 2 to 5 *Sfi*I fragments present in a particular *Sfi*I end-class pot.

As each *Sfi*I fragment is about 65 kb in length, on average, roughly 130 *Fok*I sites are anticipated per *Sfi*I fragment. Even with as many as 5 *Sfi*I fragments in a single *Sfi*I end-class pot, only 650 *Fok*I fragments would be expected. As there are 256 possible *Fok*I cohesive end sequences, any one *Fok*I end is likely to be represented only 2 to 5 times in any one *Sfi*I end-class pot. With 32 892 fragment classes defined by 4-nt indexable ends, it is anticipated that each of the (roughly) 650 *Fok*I fragments present in a *Sfi*I end-class pot will be unique in its class (i.e. will be the only fragment in the population targeted by a particular pair of indexers).

Following isolation by bead capture, each *Fok*I end-class subpot bead suspension is aliquoted into 10 standard-format multiplex indexing ligations. Amplification of multiplex indexing ligation products is performed, using the (unbiotinylated) Sfi common primer, and PCR products are analyzed by gel electrophoresis. The "known" end of each fragment in the *Fok*I end-class subpot is targeted by the biotinylated indexer by which the fragment was bead-captured; sequence information at the "unknown" cohesive end of each fragment, in addition to its length, is generated by multiplex indexing. As each *Fok*I fragment has two 4b5' cohesive ends, each fragment in an SfiI end-class pot will be multiplex-indexed twice. This provides assurance that each *Fok*I cohesive end sequence has been correctly identified on each fragment, and confirms the length of the fragment for subcontig assembly purposes.

Indexed *Fok*I fragments amplified during multiplex indexing may be isolated from agarose gel for use as templates for direct cycle sequencing. The SfiCC 4b5' NoP compound indexers used in the multiplex mixes are designed to permit directional sequencing using the SfiCC directional sequencing primer. As each *Fok*I fragment has been multiplex-indexed twice, once targeting one cohesive end sequence and once using the other, templates initiating strand elongation in both directions are available. If desired, specific amplification of any *Fok*I fragment may be performed as an additional step, circumventing the need for template purification from agarose gel. Sequencing of the S/F fragments at either end of *Sfi*I fragment contigs may be performed by directional cycle sequencing of the bead-captured S/F fragment templates, or by specific amplification and sequencing of iSTS *Fok*I fragments using Sfi P-indexers and SfiCC NoP compound indexers. Sequence data is obtained by commercial automated fluorescence-based DNA sequencing instrumentation. Sequence data for each *Fok*I fragment template is manipulated and assembled using commercially-available or proprietary bioinformatics software.

At this stage, the number of *Fok*I fragments generated from the 2 to 5 *Sfi*I fragments in each *Sfi*I end-class pot has been determined, the two 4b5' cohesive end sequences and length of each *Fok*I fragment have been defined, and the amplified indexed fragments have been used as sequencing templates for complete sequencing of all genomic DNA. Ordering of *Fok*I fragments by jigsaw assembly into subcontigs along the "backbone" of the 2 to 5 *Sfi*I fragments in a pot is possible in most cases. Using the S/F fragment that defines a particular *Sfi*I fragment within a pot as a starting point, ordering of *Fok*I fragments by matching complementary cohesive end sequences proceeds towards the opposite end of the *Sfi*I fragment. The presence of the same cohesive end sequence on multiple fragments within an *Sfi*I-pot *Fok*I fragment population complicates this approach, generating multiple putative fragment maps. Such cases are exacerbated by the lack of information regarding which *Sfi*I fragment in the *Sfi*I pot a particular *Fok*I fragment is derived from. Information from other *Sfi*I end-class pots alleviate this problem. As each *Sfi*I fragment is present in two *Sfi*I end-class pots, the subcontig assembly data for each *Sfi*I fragment, in the form of *Fok*I subpot multiplex-indexing information, is generated for each of those pots. Using the iSTS mapping data and by comparing the *Fok*I end sequence and fragment length information from each of the two end-class pots containing a particular *Sfi*I fragment, a unique jigsaw-assembly solution can generally be established for the *Fok*I subcontig for each *Sfi*I fragment. In cases where a unique subcontig assembly solution is not possible, or in instances in which greater redundancy in map construction and sequence template generation is desirable, a second iteration of subcontig mapping and sequencing using *Sfa*NI endonuclease may be performed.

5.2.3.5 Sfa*NI-based subcontig mapping and sequencing of* SfiI *fragments*

The Type IIS restriction endonuclease *Sfa*NI recognizes the sequence 5'-GGATG-3', cleaving DNA phosphodiester groups 5 bp and 9 bp away (in the 5' direction) from the recognition site. Thus, like *Fok*I, *Sfa*NI generates restriction fragments with 4b5' informative cohesive end sequences. From an indexing perspective *Sfa*NI is analogous to *Fok*I in every respect except for recognition specificity. As a result, *Sfa*NI may be used as a replacement for *Fok*I in instances in which *Fok*I recognition specificity alone is insufficient, or in which a second iteration of indexing information is required.

Repetition of the steps described for *Fok*I-based subcontig mapping and sequencing of *Sfi*I fragments in each *Sfi*I end-class pot may be performed using *Sfa*NI if desired or if necessary. The information generated by a second iteration of subcontig mapping and sequencing may be useful in instances in which map construction or sequence generation with *Fok*I alone does not permit a unique subcontig assembly solution for a particular *Sfi*I fragment. In general, a full-scale repetition across all *Sfi*I end-class pots is unnecessary. Generation of a second, *Sfa*NI-based, iSTS physical map is also likely to be unnecessary, as the *Fok*I-based iSTS physical map should provide a unique mapping solution with a resolution of about 60-70 kb.

5.2.3.6 Assembly of sequenced SfiI fragment contigs along the iSTS physical map for complete genome sequence construction

The final step required for complete sequence assembly of a bacterial genome using the IDBG approach is the alignment of the sequenced *Sfi*I-fragment contigs along the scaffold of the iSTS physical map. Each *Sfi*I-fragment contig has been identified at either end with S/F fragments of known end sequence and fragment length which form one-half of an iSTS site defined by a *Sfi*I-site-spanning *Fok*I fragment. This information is used to align each of the sequenced *Sfi*I-fragment contigs with its corresponding iSTS markers along the physical map generated previously. Each iSTS marker links two adjacent *Sfi*I-fragment contigs. When the iSTS physical map has been tiled by the full set of *Sfi*I-fragment contigs, assembly of the entire genomic sequence of the prokaryote of interest is complete.

6 Bibliography

1. Green ED, Green P: **Sequence-tagged site (STS) content mapping of human chromosomes: theoretical considerations and early experiences**. *PCR Methods Appl* 1991, 1:77-90.
2. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment**. *Genome Res* 1998, 8:175-185.
3. Bentley DR, Todd C, Collins J, Holland J, Dunham I, Hassock S, Bankier A, Giannelli F: **The development and application of automated gridding for efficient screening of yeast and bacterial ordered libraries**. *Genomics* 1992, 12:534-541.
4. Hunkapiller T, Kaiser RJ, Koop BF, Hood L: **Large-scale and automated DNA sequence determination**. *Science* 1991, 254:59-67.
5. Kern S, Hampton GM: **Direct hybridization of large-insert genomic clones on high-density gridded cDNA filter arrays**. *Biotechniques* 1997, 23:120-124.
6. Billings PR, Smith CL, Cantor CR: **New techniques for physical mapping of the human genome**. *Faseb J* 1991, 5:28-34.
7. Connelly C, McCormick MK, Shero J, Hieter P: **Polyamines eliminate an extreme size bias against transformation of large yeast artificial chromosome DNA**. *Genomics* 1991, 10:10-16.
8. Sternberg NL: **Cloning high molecular weight DNA fragments by the bacteriophage P1 system**. *Trends Genet* 1992, 8:11-16.
9. Frishman D, Mewes HW: **Genome-based structural biology**. *Prog Biophys Mol Biol* 1999, 72:1-17.
10. Hunkapiller MW: **Advances in DNA sequencing technology**. *Curr Opin Genet Dev* 1991, 1:88-92.
11. Dolinski K, Ball CA, Chervitz SA, Dwight SS, Harris MA, Roberts S, Roe T, Cherry JM, Botstein D: **Expanding yeast knowledge online**. *Yeast* 1998, 14:1453-1469.
12. Platt DM, Dix TI: **Comparison of clone-ordering algorithms used in physical mapping**. *Genomics* 1997, 40:490-492.

13. Heber S, Hoheisel J, Vingron M: **Application of bootstrap techniques to physical mapping**. *Genomics* 2000, 69:235-241.
14. Sentry LW, Kaiser K: **Progress in *Drosophila* genome manipulation.** *Transgenic Res* 1995, 4:155-162.
15. Tanksley SD, Ganal MW, Martin GB: **Chromosome landing: a paradigm for map-based gene cloning in plants with large genomes**. *Trends Genet* 1995, 11:63-68.
16. Duret L, Bucher P: **Searching for regulatory elements in human noncoding sequences**. *Curr Opin Struct Biol* 1997, 7:399-406.
17. Scherf M, Klingenhoff A, Werner T: **Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach**. *J Mol Biol* 2000, 297:599-606.
18. Toba G, Ohsako T, Miyata N, Ohtsuka T, Seong KH, Aigaki T: **The gene search system. A method for efficient detection and rapid molecular identification of genes in *Drosophila melanogaster***. *Genetics* 1999, 151:725-737.
19. Kel AE, Kondrakhin YV, Kolpakov Ph A, Kel OV, Romashenko AG, Wingender E, Milanesi L, Kolchanov NA: **Computer tool FUNSITE for analysis of eukaryotic regulatory genomic sequences**. *Proc Int Conf Intell Syst Mol Biol* 1995, 3:197-205.
20. Brazma A, Vilo J, Ukkonen E, Valtonen K: **Data mining for regulatory elements in yeast genome**. *Proc Int Conf Intell Syst Mol Biol* 1997, 5:65-74.
21. Ito T, Sakaki Y: **Toward genome-wide scanning of gene expression: a functional aspect of the Genome Project**. *Essays Biochem* 1996, 31:11-21.
22. Unrau P, Deugau KV: **Non-cloning amplification of specific DNA fragments from whole genomic DNA digests using DNA 'indexers'**. *Gene* 1994, 145:163-169.
23. Szybalski W, Kim SC, Hasan N, Podhajska AJ: **Class-IIS restriction enzymes--a review**. *Gene* 1991, 100:13-26.
24. Berger SL: **Expanding the potential of restriction endonucleases: use of hapaxoterministic enzymes**. *Anal Biochem* 1994, 222:1-8.
25. Sugisaki H, Kanazawa S: **New restriction endonucleases from *Flavobacterium okeanokoites* (*Fok*I) and *Micrococcus luteus* (*Mlu*I)**. *Gene* 1981, 16:73-78.

26. Murray NE: **Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle)**. *Microbiol Mol Biol Rev* 2000, 64:412-434.
27. Bickle TA, Kruger DH: **Biology of DNA restriction**. *Microbiol Rev* 1993, 57:434-450.
28. Kessler C, Manta V: **Specificity of restriction endonucleases and DNA modification methyltransferases a review (Edition 3)**. *Gene* 1990, 92:1-248.
29. Roberts RJ: **Restriction endonucleases**. *CRC Crit Rev Biochem* 1976, 4:123-164.
30. (NEB) NEB: *NEB Catalog 2000*. Beverly, MA: New England Biolabs; 2000.
31. Kessler C, Holtke HJ: **Specificity of restriction endonucleases and methylases--a review**. *Gene* 1986, 47:1-153.
32. Higgins NP, Cozzarelli NR: **DNA-joining enzymes: a review**. *Methods Enzymol* 1979, 68:50-71.
33. Lehman IR: **DNA ligase: structure, mechanism, and function**. *Science* 1974, 186:790-797.
34. Olivera BM, Lehman IR: **Diphosphopyridine nucleotide: a cofactor for the polynucleotide-joining enzyme from *Escherichia coli***. *Proc Natl Acad Sci U S A* 1967, 57:1700-1704.
35. Doherty AJ, Suh SW: **Structural and mechanistic conservation in DNA ligases**. *Nucleic Acids Res* 2000, 28:4051-4058.
36. Rossi R, Montecucco A, Ciarrocchi G, Biamonti G: **Functional characterization of the T4 DNA ligase: a new insight into the mechanism of action**. *Nucleic Acids Res* 1997, 25:2106-2113.
37. Khorana HG, Agarwal KL, Buchi H, Caruthers MH, Gupta NK, Kleppe K, Kumar A, Otsuka E, RajBhandary UL, Van de Sande JH, Sgaramella V, Terao T, Weber H, Yamada T: **Studies on polynucleotides. 103. Total synthesis of the structural gene for an alanine transfer ribonucleic acid from yeast**. *J Mol Biol* 1972, 72:209-217.
38. Pritchard CE, Southern EM: **Effects of base mismatches on joining of short oligodeoxynucleotides by DNA ligases**. *Nucleic Acids Res* 1997, 25:3403-3407.
39. Luo J, Bergstrom DE, Barany F: **Improving the fidelity of *Thermus thermophilus* DNA ligase**. *Nucleic Acids Res* 1996, 24:3071-3078.

40. Bailly C, Minnock A, Waring MJ: **A simple ligation assay to detect effects of drugs on the curvature/flexibility of DNA**. *FEBS Lett* 1996, 396:253-256.

41. Sgaramella V, Khorana HG: **CXII. Total synthesis of the structural gene for an alanine transfer RNA from yeast. Enzymic joining of the chemically synthesized polydeoxynucleotides to form the DNA duplex representing nucleotide sequence 1 to 20**. *J Mol Biol* 1972, 72:427-444.

42. Wiaderkiewicz R, Ruiz-Carrillo A: **Mismatch and blunt to protruding-end joining by DNA ligases**. *Nucleic Acids Res* 1987, 15:7831-7848.

43. Landegren U, Kaiser R, Sanders J, Hood L: **A ligase-mediated gene detection technique**. *Science* 1988, 241:1077-1080.

44. Brown T, Kennard O, Kneale G, Rabinovich D: **High-resolution structure of a DNA helix containing mismatched base pairs**. *Nature* 1985, 315:604-606.

45. Harada K, Orgel LE: **Unexpected substrate specificity of T4 DNA ligase revealed by *in vitro* selection**. *Nucleic Acids Res* 1993, 21:2287-2291.

46. Ferretti L, Sgaramella V: **Temperature dependence of the joining by T4 DNA ligase of termini produced by type II restriction endonucleases**. *Nucleic Acids Res* 1981, 9:85-93.

47. Harvey CL, Wright R: **Ligase joining of oligodeoxythymidylates**. *Biochemistry* 1972, 11:2667-2671.

48. Lund AH, Duch M, Pedersen FS: **Increased cloning efficiency by temperature-cycle ligation**. *Nucleic Acids Res* 1996, 24:800-801.

49. Sambrook J, Fritsch EF, Maniatis T: *Molecular Cloning: A Laboratory Manual*, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1989.

50. Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K: *Current Protocols in Molecular Biology*. New York, NY: John Wiley & Sons, Inc.; 2001.

51. Upcroft P, Healey A: **Rapid and efficient method for cloning of blunt-ended DNA fragments**. *Gene* 1987, 51:69-75.

52. Hayashi K, Nakazawa M, Ishizaki Y, Obayashi A: **Influence of monovalent cations on the activity of T4 DNA ligase in the presence of polyethylene glycol**. *Nucleic Acids Res* 1985, 13:3261-3271.

53. Mead DA, Pey NK, Herrnstadt C, Marcil RA, Smith LM: **A universal method for the direct cloning of PCR amplified nucleic acid**. *Biotechnology (N Y)* 1991, 9:657-663.

54. D'Souza CR, Deugau KV, Spencer JH: **A simplified procedure for cDNA and genomic library construction using nonpalindromic oligonucleotide adaptors**. *Biochem Cell Biol* 1989, 67:205-209.

55. Bellemare G, Potvin C, Bergeron D: **High-yield method for directional cDNA library construction**. *Gene* 1991, 98:231-235.

56. Boyd AC: **Turbo cloning: a fast, efficient method for cloning PCR products and other blunt-ended DNA fragments into plasmids**. *Nucleic Acids Res* 1993, 21:817-821.

57. Sobczak J, Duguet M: **Effect of histone H1, poly(ethyleneglycol) and DNA concentration on intermolecular and intramolecular ligation by T4 DNA ligase**. *Eur J Biochem* 1988, 175:379-385.

58. Pheiffer BH, Zimmerman SB: **Polymer-stimulated ligation: enhanced blunt- or cohesive-end ligation of DNA or deoxyribooligonucleotides by T4 DNA ligase in polymer solutions**. *Nucleic Acids Res* 1983, 11:7853-7871.

59. Hayashi K, Nakazawa M, Ishizaki Y, Hiraoka N, Obayashi A: **Regulation of inter- and intramolecular ligation with T4 DNA ligase in the presence of polyethylene glycol**. *Nucleic Acids Res* 1986, 14:7617-7631.

60. Rusche JR, Howard-Flanders P: **Hexamine cobalt chloride promotes intermolecular ligation of blunt end DNA fragments by T4 DNA ligase**. *Nucleic Acids Res* 1985, 13:1997-2008.

61. Zimmerman SB, Pheiffer BH: **Macromolecular crowding allows blunt-end ligation by DNA ligases from rat liver or *Escherichia coli***. *Proc Natl Acad Sci U S A* 1983, 80:5852-5856.

62. Poso H, Kuosmanen M: **Spermidine and spermine stimulate the activity of T4-DNA ligase**. *Biochem Biophys Res Commun* 1983, 117:217-222.

63. Sheflin LG, Fucile NW, Spaulding SW: **HMG 14 and protamine enhance ligation of linear DNA to form linear multimers: phosphorylation of HMG 14 at Ser 20 specifically inhibits intermolecular DNA ligation**. *Biochem Biophys Res Commun* 1991, 174:660-666.

64. Cimmino C, Santori F, Donini P: **Ligation of nonmatching DNA molecule ends**. *Plasmid* 1995, 34:1-10.

65. Pfeiffer P, Vielmetter W: **Joining of nonhomologous DNA double strand breaks *in vitro***. *Nucleic Acids Res* 1988, 16:907-924.

66. Wu DY, Wallace RB: **Specificity of the nick-closing activity of bacteriophage T4 DNA ligase**. *Gene* 1989, 76:245-254.

67. Ortiz T, Daza P, Pinero J, Cortes F: **T4 DNA ligase modulates chromosome damage induced by restriction endonucleases through an error-free process**. *Mutagenesis* 1995, 10:399-402.

68. Cherepanov A, Yildirim E, de Vries S: **Joining of short DNA oligonucleotides with base pair mismatches by T4 DNA ligase**. *J Biochem (Tokyo)* 2001, 129:61-68.

69. Barany F: **Genetic disease detection and DNA amplification using cloned thermostable ligase**. *Proc Natl Acad Sci U S A* 1991, 88:189-193.

70. Drabek J: **A commented dictionary of techniques for genotyping**. *Electrophoresis* 2001, 22:1024-1045.

71. Schweitzer B, Kingsmore S: **Combining nucleic acid amplification and detection**. *Curr Opin Biotechnol* 2001, 12:21-27.

72. Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW: **Yeast microarrays for genome wide parallel genetic and gene expression analysis**. *Proc Natl Acad Sci U S A* 1997, 94:13057-13062.

73. Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, Bennett HA, He YD, Dai H, Walker WL, Hughes TR, Tyers M, Boone C, Friend SH: **Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles**. *Science* 2000, 287:873-880.

74. Yanisch-Perron C, Vieira J, Messing J: **Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors**. *Gene* 1985, 33:103-119.

75. Kaczorowski T, Skowron P, Podhajska AJ: **Purification and characterization of the *Fok*I restriction endonuclease**. *Gene* 1989, 80:209-216.

76. Wah DA, Hirsch JA, Dorner LF, Schildkraut I, Aggarwal AK: **Structure of the multimodular endonuclease *Fok*I bound to DNA**. *Nature* 1997, 388:97-100.

77. Skowron P, Kaczorowski T, Tucholski J, Podhajska AJ: **Atypical DNA-binding properties of class-IIS restriction endonucleases: evidence for recognition of the cognate sequence by a *Fok*I monomer**. *Gene* 1993, 125:1-10.

78. Li L, Wu LP, Chandrasegaran S: **Functional domains in *Fok*I restriction endonuclease**. *Proc Natl Acad Sci U S A* 1992, 89:4275-4279.

79. Li L, Wu LP, Clarke R, Chandrasegaran S: **C-terminal deletion mutants of the *Fok*I restriction endonuclease**. *Gene* 1993, 133:79-84.

80. Kim YG, Li L, Chandrasegaran S: **Insertion and deletion mutants of *Fok*I restriction endonuclease**. *J Biol Chem* 1994, 269:31978-31982.

81. Waugh DS, Sauer RT: **A novel class of *Fok*I restriction endonuclease mutants that cleave hemi- methylated substrates**. *J Biol Chem* 1994, 269:12298-12303.

82. Yonezawa A, Sugiura Y: **DNA binding mode of class-IIS restriction endonuclease *Fok*I revealed by DNA footprinting analysis**. *Biochim Biophys Acta* 1994, 1219:369-379.

83. Meunier JR, Grimont PA: **Factors affecting reproducibility of random amplified polymorphic DNA fingerprinting**. *Res Microbiol* 1993, 144:373-379.

84. Lantz PG, Abu al-Soud W, Knutsson R, Hahn-Hagerdal B, Radstrom P: **Biotechnical use of polymerase chain reaction for microbiological analysis of biological samples**. *Biotechnol Annu Rev* 2000, 5:87-130.

85. Haworth R, Pilling AM: **The PCR assay in the preclinical safety evaluation of nucleic acid medicines**. *Hum Exp Toxicol* 2000, 19:267-276.

86. Hohoff C, Brinkmann B: **Human identity testing with PCR-based systems**. *Mol Biotechnol* 1999, 13:123-136.

87. Freeman WM, Walker SJ, Vrana KE: **Quantitative RT-PCR: pitfalls and potential**. *Biotechniques* 1999, 26:112-122, 124-115.

88. Orlando C, Pinzani P, Pazzagli M: **Developments in quantitative PCR**. *Clin Chem Lab Med* 1998, 36:255-269.

89. Power EG: **RAPD typing in microbiology--a technical review**. *J Hosp Infect* 1996, 34:247-265.

90. Komminoth P, Long AA: ***In-situ* polymerase chain reaction. An overview of methods, applications and limitations of a new molecular technique**. *Virchows Arch B Cell Pathol Incl Mol Pathol* 1993, 64:67-73.

91. Chien A, Edgar DB, Trela JM: **Deoxyribonucleic acid polymerase from the extreme thermophile *Thermus aquaticus***. *J Bacteriol* 1976, 127:1550-1557.

92. Lawyer FC, Stoffel S, Saiki RK, Myambo K, Drummond R, Gelfand DH: **Isolation, characterization, and expression in *Escherichia coli* of the DNA polymerase gene from *Thermus aquaticus*.** *J Biol Chem* 1989, 264:6427-6437.

93. Tindall KR, Kunkel TA: **Fidelity of DNA synthesis by the *Thermus aquaticus* DNA polymerase.** *Biochemistry* 1988, 27:6008-6013.

94. Petruska J, Goodman MF, Boosalis MS, Sowers LC, Cheong C, Tinoco I, Jr.: **Comparison between DNA melting thermodynamics and DNA polymerase fidelity.** *Proc Natl Acad Sci U S A* 1988, 85:6252-6256.

95. Rychlik W, Spencer WJ, Rhoads RE: **Optimization of the annealing temperature for DNA amplification *in vitro*.** *Nucleic Acids Res* 1990, 18:6409-6412.

96. Korolev S, Nayal M, Barnes WM, Di Cera E, Waksman G: **Crystal structure of the large fragment of *Thermus aquaticus* DNA polymerase I at 2.5-A resolution: structural basis for thermostability.** *Proc Natl Acad Sci U S A* 1995, 92:9264-9268.

97. Kim Y, Eom SH, Wang J, Lee DS, Suh SW, Steitz TA: **Crystal structure of *Thermus aquaticus* DNA polymerase.** *Nature* 1995, 376:612-616.

98. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA: **Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase.** *Science* 1988, 239:487-491.

99. Kricker MC, Tindall KR: **Direct sequencing of bacteriophage T4 DNA with a thermostable DNA polymerase.** *Gene* 1989, 85:199-204.

100. Timblin C, Battey J, Kuehl WM: **Application for PCR technology to subtractive cDNA cloning: identification of genes expressed specifically in murine plasmacytoma cells.** *Nucleic Acids Res* 1990, 18:1587-1593.

101. Jones MD: **Reverse transcription of mRNA by *Thermus aquaticus* DNA polymerase followed by polymerase chain reaction amplification.** *Methods Enzymol* 1993, 218:413-419.

102. Barnes WM: **PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates.** *Proc Natl Acad Sci U S A* 1994, 91:2216-2220.

103. Cairns MJ, Murray V: **Dideoxy genomic sequencing of a single-copy mammalian gene using more than two hundred cycles of linear amplification**. *Biotechniques* 1994, 17:910-914.

104. Rychlik W: **Selection of primers for polymerase chain reaction**. *Mol Biotechnol* 1995, 3:129-134.

105. Lundberg KS, Shoemaker DD, Adams MW, Short JM, Sorge JA, Mathur EJ: **High-fidelity amplification using a thermostable DNA polymerase isolated from *Pyrococcus furiosus***. *Gene* 1991, 108:1-6.

106. Brail L, Fan E, Levin DB, Logan DM: **Improved polymerase fidelity in PCR-SSCPA**. *Mutat Res* 1993, 303:171-175.

107. Borns M, Cline J, Allen R, Hogrefe H: **Comparing PCR performance and fidelity of commercial *Pfu* DNA polymerases**. In: *Strategies Newsletter*, vol. 12. Stratagene Inc. 1999.

108. Li Q, Ownby CL: **A rapid method for extraction of DNA from agarose gels using a syringe**. *Biotechniques* 1993, 15:976-978.

109. Roos P, Puyang X: Personal communication. 1998.

110. Bay SJ: *The construction and evaluation of a multiple capillary DNA sequencer*. Edmonton, AB: University of Alberta; 1998.

111. Kita K, Kotani H, Hiraoka N, Nakamura T, Yonaha K: **Overproduction and crystallization of *Fok*I restriction endonuclease**. *Nucleic Acids Res* 1989, 17:8741-8753.

112. Nakagawa Y, Sakane T, Yokota A: **Emendation of the genus *Planococcus* and transfer of *Flavobacterium okeanokoites* Zobell and Upham 1944 to the genus *Planococcus* as *Planococcus okeanokoites* comb. nov**. *Int J Syst Bacteriol* 1996, 46:866-870.

113. Sheridan PP, Brenchley JE: **Characterization of a salt-tolerant family 42 beta-galactosidase from a psychrophilic antarctic *Planococcus* isolate**. *Appl Environ Microbiol* 2000, 66:2438-2444.

114. Junge K, Gosink JJ, Hoppe HG, Staley JT: ***Arthrobacter*, *Brachybacterium* and *Planococcus* isolates identified from antarctic sea ice brine. Description of *Planococcus mcmeekinii*, sp. nov.** *Syst Appl Microbiol* 1998, 21:306-314.

115. Chou Q, Russell M, Birch DE, Raymond J, Bloch W: **Prevention of pre-PCR mis-priming and primer dimerization improves low-copy-number amplifications**. *Nucleic Acids Res* 1992, 20:1717-1723.

116. Mullis KB: **The polymerase chain reaction in an anemic mode: how to avoid cold oligodeoxyribonuclear fusion**. *PCR Methods Appl* 1991, 1:1-4.

117. Sharkey DJ, Scalice ER, Christy KG, Jr., Atwood SM, Daiss JL: **Antibodies as thermolabile switches: high temperature triggering for the polymerase chain reaction**. *Biotechnology (N Y)* 1994, 12:506-509.

118. Chumakov KM: **Reverse transcriptase can inhibit PCR and stimulate primer-dimer formation**. *PCR Methods Appl* 1994, 4:62-64.

119. Don RH, Cox PT, Wainwright BJ, Baker K, Mattick JS: **'Touchdown' PCR to circumvent spurious priming during gene amplification**. *Nucleic Acids Res* 1991, 19:4008.

120. Blair P, Ramanujam R, Burdick BA: **Wax-embedded PCR reagents**. *PCR Methods Appl* 1994, 4:191-194.

121. Higuchi R, Krummel B, Saiki RK: **A general method of in vitro preparation and specific mutagenesis of DNA fragments: study of protein and DNA interactions**. *Nucleic Acids Res* 1988, 16:7351-7367.

122. Brownie J, Shawcross S, Theaker J, Whitcombe D, Ferrie R, Newton C, Little S: **The elimination of primer-dimer accumulation in PCR**. *Nucleic Acids Res* 1997, 25:3235-3241.

123. Kellogg DE, Rybalkin I, Chen S, Mukhamedova N, Vlasik T, Siebert PD, Chenchik A: **TaqStart Antibody: "hot start" PCR facilitated by a neutralizing monoclonal antibody directed against *Taq* DNA polymerase**. *Biotechniques* 1994, 16:1134-1137.

124. Scalice ER, Sharkey DJ, Daiss JL: **Monoclonal antibodies prepared against the DNA polymerase from *Thermus aquaticus* are potent inhibitors of enzyme activity**. *J Immunol Methods* 1994, 172:147-163.

125. Ogata N, Miura T: **Creation of genetic information by DNA polymerase of the thermophilic bacterium *Thermus thermophilus***. *Nucleic Acids Res* 1998, 26:4657-4661.

126. Ogata N, Miura T: **Creation of genetic information by DNA polymerase of the archaeon *Thermococcus litoralis*: influences of temperature and ionic strength**. *Nucleic Acids Res* 1998, 26:4652-4656.

127. Carlisle SM, Deugau KV, Unrau P: Personal communication. 1997.

128. Chang ACY, Heyneker HL: Personal communication. 1997.

129. Ayyadevara S, Thaden JJ, Shmookler Reis RJ: **Discrimination of primer 3'-nucleotide mismatch by *Taq* DNA polymerase during polymerase chain reaction**. *Anal Biochem* 2000, 284:11-18.

130. Hager J: Personal communication. 2001.

131. Unrau P, Deugau KV: Personal communication. 1998.

132. Tenover FC, Arbeit R, Archer G, Biddle J, Byrne S, Goering R, Hancock G, Hebert GA, Hill B, Hollis R, *et al.*: **Comparison of traditional and molecular methods of typing isolates of *Staphylococcus aureus***. *J Clin Microbiol* 1994, 32:407-415.

133. On SL: **Identification methods for campylobacters, helicobacters, and related organisms**. *Clin Microbiol Rev* 1996, 9:405-422.

134. Hawkey PM: **Principles of molecular typing: a guide to the letters**. *J Hosp Infect* 1999, 43 Suppl:S77-83.

135. van Belkum A, Struelens M, de Visser A, Verbrugh H, Tibayrenc M: **Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology**. *Clin Microbiol Rev* 2001, 14:547-560.

136. Matar GM, Swaminathan B, Hunter SB, Slater LN, Welch DF: **Polymerase chain reaction-based restriction fragment length polymorphism analysis of a fragment of the ribosomal operon from *Rochalimaea* species for subtyping**. *J Clin Microbiol* 1993, 31:1730-1734.

137. Tenover FC, Arbeit RD, Goering RV: **How to select and interpret molecular strain typing methods for epidemiological studies of bacterial infections: a review for healthcare epidemiologists**. *Infect Control Hosp Epidemiol* 1997, 18:426-439.

138. Stull TL, LiPuma JJ, Edlind TD: **A broad-spectrum probe for molecular epidemiology of bacteria: ribosomal RNA**. *J Infect Dis* 1988, 157:280-286.

139. Schwartz DC, Cantor CR: **Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis**. *Cell* 1984, 37:67-75.

140. Welsh J, McClelland M: **Fingerprinting genomes using PCR with arbitrary primers**. *Nucleic Acids Res* 1990, 18:7213-7218.

141. Williams JG, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV: **DNA polymorphisms amplified by arbitrary primers are useful as genetic markers**. *Nucleic Acids Res* 1990, 18:6531-6535.

142. Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, et al.: **AFLP: a new technique for DNA fingerprinting**. *Nucleic Acids Res* 1995, 23:4407-4414.

143. Janssen P, Coopman R, Huys G, Swings J, Bleeker M, Vos P, Zabeau M, Kersters K: **Evaluation of the DNA fingerprinting method AFLP as an new tool in bacterial taxonomy**. *Microbiology* 1996, 142 (Pt 7):1881-1893.

144. Duim B, Ang CW, van Belkum A, Rigter A, van Leeuwen NW, Endtz HP, Wagenaar JA: **Amplified fragment length polymorphism analysis of *Campylobacter jejuni* strains isolated from chickens and from patients with gastroenteritis or Guillain-Barre or Miller Fisher syndrome**. *Appl Environ Microbiol* 2000, 66:3917-3923.

145. Cleary PP, Kaplan EL, Livdahl C, Skjold S: **DNA fingerprints of *Streptococcus pyogenes* are M type specific**. *J Infect Dis* 1988, 158:1317-1323.

146. Fitzgerald C, Owen RJ, Stanley J: **Comprehensive ribotyping scheme for heat-stable serotypes of *Campylobacter jejuni***. *J Clin Microbiol* 1996, 34:265-269.

147. Arthur M, Arbeit RD, Kim C, Beltran P, Crowe H, Steinbach S, Campanelli C, Wilson RA, Selander RK, Goldstein R: **Restriction fragment length polymorphisms among uropathogenic *Escherichia coli* isolates: pap-related sequences compared with *rrn* operons**. *Infect Immun* 1990, 58:471-479.

148. Wassenaar TM, Newell DG: **Genotyping of *Campylobacter* spp**. *Appl Environ Microbiol* 2000, 66:1-9.

149. Nair S, Schreiber E, Thong KL, Pang T, Altwegg M: **Genotypic characterization of *Salmonella typhi* by amplified fragment length polymorphism fingerprinting provides increased discrimination as compared to pulsed-field gel electrophoresis and ribotyping**. *J Microbiol Methods* 2000, 41:35-43.

150. Arnold C, Metherell L, Clewley JP, Stanley J: **Predictive modelling of fluorescent AFLP: a new approach to the molecular epidemiology of *E. coli***. *Res Microbiol* 1999, 150:33-44.

151. Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y: **The complete genome sequence of *Escherichia coli* K-12**. *Science* 1997, 277:1453-1474.

152. Gibson TG: *Studies on the Epstein-Barr virus genome*. Cambridge, UK: University of Cambridge; 1984.

153. Palmer BR, Marinus MG: **The *dam* and *dcm* strains of *Escherichia coli*--a review**. *Gene* 1994, 143:1-12.

154. Yamamoto Y, Aiba H, Baba T, Hayashi K, Inada T, Isono K, Itoh T, Kimura S, Kitagawa M, Makino K, Miki T, Mitsuhashi N, Mizobuchi K, Mori H, Nakade S, Nakamura Y, Nashimoto H, Oshima T, Oyama S, Saito N, Sampei G, Satoh Y, Sivasundaram S, Tagami H, Horiuchi T, et al.: **Construction of a contiguous 874-kb sequence of the *Escherichia coli* K-12 genome corresponding to the 50.0-68.8 min on the linkage map and analysis of its sequence features (supplement)**. *DNA Res* 1997, 4:169-178.

155. *E. coli* MG1655, complete genomic sequence. http://www.genome.wisc.edu/pub/sequence/ecolim52.seq.

156. Barany F, Gelfand DH: **Cloning, overexpression and nucleotide sequence of a thermostable DNA ligase-encoding gene**. *Gene* 1991, 109:1-11.

157. Zimmerman SB, Little JW, Oshinsky CK, Gellert M: **Enzymatic joining of DNA strands: a novel reaction of diphosphopyridine nucleotide**. *Proc Natl Acad Sci U S A* 1967, 57:1841-1848.

158. Takahashi M, Yamaguchi E, Uchida T: **Thermophilic DNA ligase. Purification and properties of the enzyme from *Thermus thermophilus* HB8**. *J Biol Chem* 1984, 259:10041-10047.

159. Kato K: **Description of the entire mRNA population by a 3' end cDNA fragment generated by class IIS restriction enzymes**. *Nucleic Acids Res* 1995, 23:3685-3690.

160. Protein list for *Escherichia coli* K12, complete genome. http://www.ncbi.nlm.nih.gov:80/cgi-bin/Entrez/altik?gi=115&db=Genome.

161. Boyce JM: **Increasing prevalence of methicillin-resistant *Staphylococcus aureus* in the United States**. *Infect Control Hosp Epidemiol* 1990, 11:639-642.

162. Saruta K, Matsunaga T, Kono M, Hoshina S, Ikawa S, Sakai O, Machida K: **Rapid identification and typing of *Staphylococcus aureus* by nested PCR**

amplified ribosomal DNA spacer region. *FEMS Microbiol Lett* 1997, 146:271-278.

163. Wilton J, Jung K, Vedin I, Aronsson B, Flock JI: **Comparative evaluation of a new molecular method for typing *Staphylococcus epidermidis*.** *Eur J Clin Microbiol Infect Dis* 1992, 11:515-521.

164. Kleeman KT, Bannerman TL, Kloos WE: **Species distribution of coagulase-negative staphylococcal isolates at a community hospital and implications for selection of staphylococcal identification procedures.** *J Clin Microbiol* 1993, 31:1318-1321.

165. Matar GM, Koehler JE, Malcolm G, Lambert-Fair MA, Tappero J, Hunter SB, Swaminathan B: **Identification of *Bartonella* species directly in clinical specimens by PCR-restriction fragment length polymorphism analysis of a 16S rRNA gene fragment.** *J Clin Microbiol* 1999, 37:4045-4047.

166. Ochman H, Jones IB: **Evolutionary dynamics of full genome content in *Escherichia coli*.** *Embo J* 2000, 19:6637-6643.

167. Salama N, Guillemin K, McDaniel TK, Sherlock G, Tompkins L, Falkow S: **A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains.** *Proc Natl Acad Sci U S A* 2000, 97:14668-14673.

168. Fitzgerald JR, Sturdevant DE, Mackie SM, Gill SR, Musser JM: **Evolutionary genomics of *Staphylococcus aureus*: Insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic.** *Proc Natl Acad Sci U S A* 2001, 98:8821-8826.

169. Musser JM: **Molecular population genetic analysis of emerged bacterial pathogens: selected insights.** *Emerg Infect Dis* 1996, 2:1-17.

170. Kreiswirth B, Kornblum J, Arbeit RD, Eisner W, Maslow JN, McGeer A, Low DE, Novick RP: **Evidence for a clonal origin of methicillin resistance in *Staphylococcus aureus*.** *Science* 1993, 259:227-230.

171. Franklin RB, Taylor DR, Mills AL: **Characterization of microbial communities using randomly amplified polymorphic DNA (RAPD).** *J Microbiol Methods* 1999, 35:225-235.

172. Amann R, Ludwig W: **Ribosomal RNA-targeted nucleic acid probes for studies in microbial ecology.** *FEMS Microbiol Rev* 2000, 24:555-565.

173. Holben WE, Harris D: **DNA-based monitoring of total bacterial community structure in environmental samples.** *Mol Ecol* 1995, 4:627-631.

174. Holben WE: **Isolation and purification of bacterial community DNA from environmental samples**. In: *Manual of Environmental Microbiology* Edited by Hurst CJ, Knudsen GR, McIerney MJ, Stetzenback LD, Walter MV. pp. 431-444. Washington, DC: American Society for Microbiology; 1997.

175. Tiedje JM, Asuming-Brempong S, Nusslein K, Marsh TL, Flynn SJ: **Opening the black box of soil microbial diversity**. *Applied Soil Ecology* 1999, 13:109-122.

176. Ranjard L, Poly F, Nazaret S: **Monitoring complex bacterial communities using culture-independent molecular techniques: application to soil environment**. *Research in Microbiology* 2000, 151:167-177.

177. Woese CR: **Bacterial evolution**. *Microbiol Rev* 1987, 51:221-271.

178. Bruce K: **Analysis of *mer* gene subclasses within bacterial communities in soils and sediments resolved by fluorescent-PCR-restriction fragment length polymorphism profiling**. *Appl. Environ. Microbiol.* 1997, 63:4914-4919.

179. Marsh TL: **Terminal restriction fragment length polymorphism (T-RFLP): an emerging method for characterizing diversity among homologous populations of amplification products**. *Curr Opin Microbiol* 1999, 2:323-327.

180. Maidak BL, Cole JR, Parker CT, Jr., Garrity GM, Larsen N, Li B, Lilburn TG, McCaughey MJ, Olsen GJ, Overbeek R, Pramanik S, Schmidt TM, Tiedje JM, Woese CR: **A new version of the RDP (Ribosomal Database Project)**. *Nucleic Acids Res* 1999, 27:171-173.

181. Van de Peer Y, Robbrecht E, de Hoog S, Caers A, De Rijk P, De Wachter R: **Database on the structure of small subunit ribosomal RNA**. *Nucleic Acids Res* 1999, 27:179-183.

182. Dunbar J, Ticknor LO, Kuske CR: **Phylogenetic specificity and reproducibility and new method for analysis of terminal restriction fragment profiles of 16S rRNA genes from bacterial communities**. *Appl Environ Microbiol* 2001, 67:190-197.

183. Moeseneder MM, Arrieta JM, Muyzer G, Winter C, Herndl GJ: **Optimization of terminal-restriction fragment length polymorphism analysis for complex marine bacterioplankton communities and comparison with denaturing gradient gel electrophoresis**. *Appl Environ Microbiol* 1999, 65:3518-3525.

184. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW: **A genome-wide transcriptional analysis of the mitotic cell cycle.** *Mol Cell* 1998, 2:65-73.

185. Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW: **Gene expression profiles in normal and cancer cells**. *Science* 1997, 276:1268-1272.

186. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA: **Dissecting the regulatory circuitry of a eukaryotic genome.** *Cell* 1998, 95:717-728.

187. Zhao LP, Prentice R, Breeden L: **Statistical modeling of large microarray data sets to identify stimulus-response profiles.** *Proc Natl Acad Sci U S A* 2001, 98:5631-5636.

188. Lockhart DJ, Winzeler EA: **Genomics, gene expression and DNA arrays.** *Nature* 2000, 405:827-836.

189. Liang P, Pardee AB: **Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction.** *Science* 1992, 257:967-971.

190. Mou L, Miller H, Li J, Wang E, Chalifour L: **Improvements to the differential display method for gene analysis.** *Biochem Biophys Res Commun* 1994, 199:564-569.

191. Bertioli DJ, Schlichter UH, Adams MJ, Burrows PR, Steinbiss HH, Antoniw JF: **An analysis of differential display shows a strong bias towards high copy number mRNAs.** *Nucleic Acids Res* 1995, 23:4520-4523.

192. Liang P, Averboukh L, Pardee AB: **Distribution and cloning of eukaryotic mRNAs by means of differential display: refinements and optimization.** *Nucleic Acids Res* 1993, 21:3269-3275.

193. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, 270:484-487.

194. Brady G: **Expression profiling of single mammalian cells--small is beautiful.** *Yeast* 2000, 17:211-217.

195. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, Jr., Hieter P, Vogelstein B, Kinzler KW: **Characterization of the yeast transcriptome.** *Cell* 1997, 88:243-251.

196. Yamamoto M, Wakatsuki T, Hada A, Ryo A: **Use of serial analysis of gene expression (SAGE) technology**. *J Immunol Methods* 2001, 250:45-66.

197. Chen JJ, Rowley JD, Wang SM: **Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification**. *Proc Natl Acad Sci U S A* 2000, 97:349-353.

198. SAGEmap at National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/SAGE.

199. Stollberg J, Urschitz J, Urban Z, Boyd CD: **A quantitative evaluation of SAGE**. *Genome Res* 2000, 10:1241-1248.

200. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP: **Light-generated oligonucleotide arrays for rapid DNA sequence analysis**. *Proc Natl Acad Sci U S A* 1994, 91:5022-5026.

201. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale**. *Science* 1997, 278:680-686.

202. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: **Expression profiling using cDNA microarrays**. *Nat Genet* 1999, 21:10-14.

203. Lee ML, Kuo FC, Whitmore GA, Sklar J: **Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations**. *Proc Natl Acad Sci U S A* 2000, 97:9834-9839.

204. Afshari CA, Nuwaysir EF, Barrett JC: **Application of complementary DNA microarray technology to carcinogen identification, toxicology, and drug safety evaluation**. *Cancer Res* 1999, 59:4759-4760.

205. Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA: **Microarrays and toxicology: the advent of toxicogenomics**. *Mol Carcinog* 1999, 24:153-159.

206. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN: **A gene expression database for the molecular pharmacology of cancer**. *Nat Genet* 2000, 24:236-244.

207. Khan J, Saal LH, Bittner ML, Chen Y, Trent JM, Meltzer PS: **Expression profiling in cancer using cDNA microarrays**. *Electrophoresis* 1999, 20:223-229.

208. DeRisi J, van den Hazel B, Marc P, Balzi E, Brown P, Jacq C, Goffeau A: **Genome microarray analysis of transcriptional activation in multidrug resistance yeast mutants**. *FEBS Lett* 2000, 470:156-160.

209. Yoshikawa T, Nagasugi Y, Azuma T, Kato M, Sugano S, Hashimoto K, Masuho Y, Muramatsu M, Seki N: **Isolation of novel mouse genes differentially expressed in brain using cDNA microarray**. *Biochem Biophys Res Commun* 2000, 275:532-537.

210. Richmond CS, Glasner JD, Mau R, Jin H, Blattner FR: **Genome-wide expression profiling in *Escherichia coli* K-12**. *Nucleic Acids Res* 1999, 27:3821-3835.

211. Lucchini S, Thompson A, Hinton JC: **Microarrays for microbiologists.** *Microbiology* 2001, 147:1403-1414.

212. Schenk PM, Kazan K, Wilson I, Anderson JP, Richmond T, Somerville SC, Manners JM: **Coordinated plant defense responses in *Arabidopsis* revealed by microarray analysis**. *Proc Natl Acad Sci U S A* 2000, 97:11655-11660.

213. Anonymous: **Getting hip to the chip**. *Nature Genetics* 1998, 18:195-197.

214. Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI: **A sampling of the yeast proteome**. *Mol Cell Biol* 1999, 19:7357-7368.

215. Kato K: **Adaptor-tagged competitive PCR: a novel method for measuring relative gene expression**. *Nucleic Acids Res* 1997, 25:4694-4696.

216. Matoba R, Kato K, Kurooka C, Maruyama C, Sakakibara Y, Matsubara K: **Correlation between gene functions and developmental expression patterns in the mouse cerebellum**. *Eur J Neurosci* 2000, 12:1357-1371.

217. Matoba R, Kato K, Saito S, Kurooka C, Maruyama C, Sakakibara Y, Matsubara K: **Gene expression in mouse cerebellum during its development**. *Gene* 2000, 241:125-131.

218. Matoba R, Saito S, Ueno N, Maruyama C, Matsubara K, Kato K: **Gene expression profiling of mouse postnatal cerebellar development**. *Physiol Genomics* 2000, 4:155-164.

219. Shaw-Smith CJ, Coffey AJ, Leversha M, Freeman TC, Bentley DR, Walters JR: **Characterisation of a novel murine intestinal serine protease, DISP**. *Biochim Biophys Acta* 2000, 1490:131-136.

220. Mahadeva H, Starkey MP, Sheikh FN, Mundy CR, Samani NJ: **A simple and efficient method for the isolation of differentially expressed genes**. *J Mol Biol* 1998, 284:1391-1398.
221. Ryan M, Starkey M, Faull R, Emson P, Bahn S: **Indexing-based differential display - studies on post-mortem Alzheimer's brains**. *Brain Res Mol Brain Res* 2001, 88:199-202.
222. Kato K: **RNA fingerprinting by molecular indexing**. *Nucleic Acids Res* 1996, 24:394-395.
223. Shaw-Smith CJ, Coffey AJ, Huckle E, Durham J, Campbell EA, Freeman TC, Walters JR, Bentley DR: **Improved method for detecting differentially expressed genes using cDNA indexing**. *Biotechniques* 2000, 28:958-964.
224. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: **Life with 6000 genes**. *Science* 1996, 274:546, 563-547.
225. Mewes HW, Albermann K, Bahr M, Frishman D, Gleissner A, Hani J, Heumann K, Kleine K, Maierl A, Oliver SG, Pfeiffer F, Zollner A: **Overview of the yeast genome**. *Nature* 1997, 387:7-65.
226. Marc P, Devaux F, Jacq C: **yMGV: a database for visualization and data mining of published genome-wide yeast expression data**. *Nucleic Acids Res* 2001, 29:E63-63.
227. Thieffry D: **From global expression data to gene networks**. *Bioessays* 1999, 21:895-899.
228. Lewin B: *Genes VII*. Oxford, UK: Oxford University Press; 2000.
229. Tavazoie S, Church GM: **Quantitative whole-genome analysis of DNA-protein interactions by *in vivo* methylase protection in *E. coli***. *Nat Biotechnol* 1998, 16:566-571.
230. Yeast Microarray Global Viewer. http://www.transcriptome.ens.fr/ymgv.
231. Dambrowitz KA: *Domain Function and Regulation of Ste12p*. Vancouver, BC: University of British Columbia; 2001.
232. Tan W: *Rapid sizing of DNA and analysis of single cells using capillary electrophoresis*. Edmonton, AB: University of Alberta; 2000.

233. Kopecka M, Gabriel M: **The aberrant positioning of nuclei and the microtubular cytoskeleton in *Saccharomyces cerevisiae* due to improper actin function**. *Microbiology* 1998, 144 (Pt 7):1783-1797.

234. Bardwell L, Cook JG, Inouye CJ, Thorner J: **Signal propagation and regulation in the mating pheromone response pathway of the yeast *Saccharomyces cerevisiae***. *Dev Biol* 1994, 166:363-379.

235. Dorer R, Boone C, Kimbrough T, Kim J, Hartwell LH: **Genetic analysis of default mating behavior in *Saccharomyces cerevisiae***. *Genetics* 1997, 146:39-55.

236. Hagen DC, McCaffrey G, Sprague GF, Jr.: **Pheromone response elements are necessary and sufficient for basal and pheromone-induced transcription of the FUS1 gene of *Saccharomyces cerevisiae***. *Mol Cell Biol* 1991, 11:2952-2961.

237. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA: **Genome-wide location and function of DNA binding proteins**. *Science* 2000, 290:2306-2309.

238. Herrick D, Parker R, Jacobson A: **Identification and comparison of stable and unstable mRNAs in *Saccharomyces cerevisiae***. *Mol Cell Biol* 1990, 10:2269-2284.

239. Heiman MG, Walter P: **Prm1p, a pheromone-regulated multispanning membrane protein, facilitates plasma membrane fusion during yeast mating**. *J Cell Biol* 2000, 151:719-730.

240. Zanolari B, Riezman H: **Quantitation of alpha-factor internalization and response during the *Saccharomyces cerevisiae* cell cycle**. *Mol Cell Biol* 1991, 11:5251-5258.

241. Stefan CJ, Overton MC, Blumer KJ: **Mechanisms governing the activation and trafficking of yeast G protein-coupled receptors**. *Mol Biol Cell* 1998, 9:885-899.

242. Hartwell LH: **Mutants of *Saccharomyces cerevisiae* unresponsive to cell division control by polypeptide mating hormone**. *J Cell Biol* 1980, 85:811-822.

243. Kirkman-Correia C, Stroke IL, Fields S: **Functional domains of the yeast STE12 protein, a pheromone-responsive transcriptional activator**. *Mol Cell Biol* 1993, 13:3765-3772.

244. Adams A, Gottschling DE, Kaiser CA, Stearns T: *Methods in Yeast Genetics: A Cold Spring Harbor Laboratory Course Manual*, 1997 ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1998.

245. Lund V, Schmid R, Rickwood D, Hornes E: **Assessment of methods for covalent binding of nucleic acids to magnetic beads, Dynabeads, and the characteristics of the bound nucleic acids in hybridization reactions.** *Nucleic Acids Res* 1988, 16:10861-10880.

246. Ianakiev P: Personal communication. 2000.

247. ORF - FASTA - Saccharomyces Genome Database. ftp://genome-ftp.stanford.edu/pub/yeast/yeast_ORFs/orf_coding.fasta.Z.

248. Bhat PJ, Hopper JE: **Overproduction of the GAL1 or GAL3 protein causes galactose-independent activation of the GAL4 protein: evidence for a new model of induction for the yeast *GAL/MEL* regulon**. *Mol Cell Biol* 1992, 12:2701-2707.

249. Roth FP, Hughes JD, Estep PW, Church GM: **Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation**. *Nat Biotechnol* 1998, 16:939-945.

250. Jelinsky SA, Estep P, Church GM, Samson LD: **Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells: *Rpn4* links base excision repair with proteasomes**. *Mol Cell Biol* 2000, 20:8157-8167.

251. Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, Lee TI, True HL, Lander ES, Young RA: **Remodeling of yeast genome expression in response to environmental changes**. *Mol Biol Cell* 2001, 12:323-337.

252. YPD *Saccharomyces cerevisiae* Proteome Database. http://www.proteome.com/databases/YPD.

253. Hodges PE, McKee AH, Davis BP, Payne WE, Garrels JI: **The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data**. *Nucleic Acids Res* 1999, 27:69-73.

254. Iyer V, Struhl K: **Absolute mRNA levels and transcriptional initiation rates in *Saccharomyces cerevisiae***. *Proc Natl Acad Sci U S A* 1996, 93:5208-5212.

255. Wieland J, Nitsche AM, Strayle J, Steiner H, Rudolph HK: **The *PMR2* gene cluster encodes functionally distinct isoforms of a putative Na+ pump in the yeast plasma membrane.** *Embo J* 1995, 14:3870-3882.

256. Blomberg A: **Metabolic surprises in *Saccharomyces cerevisiae* during adaptation to saline conditions: questions, some answers and a model.** *FEMS Microbiol Lett* 2000, 182:1-8.

257. Posas F, Chambers JR, Heyman JA, Hoeffler JP, de Nadal E, Arino J: **The transcriptional response of yeast to saline stress.** *J Biol Chem* 2000, 275:17249-17255.

258. Yale J, Bohnert HJ: **Transcript expression in *Saccharomyces cerevisiae* at high salinity.** *J Biol Chem* 2001, 276:15996-16007.

259. Rep M, Krantz M, Thevelein JM, Hohmann S: **The transcriptional response of *Saccharomyces* cerevisiae to osmotic shock. Hot1p and Msn2p/Msn4p are required for the induction of subsets of high osmolarity glycerol pathway-dependent genes.** *J Biol Chem* 2000, 275:8290-8300.

260. Schulze A, Downward J: **Navigating gene expression using microarrays - a technology review.** *Nat Cell Biol* 2001, 3:E190-195.

261. Hollon T: **Comparing microarray data: What technology is needed?** *J Natl Cancer Inst* 2001, 93:1126-1127.

262. Coghlan A, Wolfe KH: **Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*.** *Yeast* 2000, 16:1131-1145.

263. Kao CM: **Functional genomic technologies: creating new paradigms for fundamental and applied biology.** *Biotechnol Prog* 1999, 15:304-311.

264. Mathieu-Daude F, Welsh J, Vogt T, McClelland M: **DNA rehybridization during PCR: the 'C_ot effect' and its consequences.** *Nucleic Acids Res* 1996, 24:2080-2086.

265. Innis MA, Gelfand DH: **Optimization of PCRs.** In: *PCR Protocols: A Guide to Methods and Applications* Edited by Innis MA, Gelfand DH, Sninsky JJ, White TJ. San Diego, CA: Academic Press, Inc.; 1990.

266. Suzuki M, Giovannoni S: **Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR.** *Appl. Environ. Microbiol.* 1996, 62:625-630.

267. Schwabe H, Stein U, Walther W: **High-copy cDNA amplification of minimal total RNA quantities for gene expression analyses**. *Mol Biotechnol* 2000, 14:165-172.

268. Shaw D: Personal communication. 1999.

269. Lander ES: **The new genomics: global views of biology**. *Science* 1996, 274:536-539.

270. Brenner S, Livak KJ: **DNA fingerprinting by sampled sequencing**. *Proc Natl Acad Sci U S A* 1989, 86:8902-8906.

271. Carlisle SM, Dambrowitz CJ, Unrau P, Deugau KV: **DNA Indexing: multiplex coding for efficient identification of Class-IIS restriction fragment ends**. In: *Human Genome Workshop*; 1994; Washington, DC.

272. Unrau P, Deugau KV: **A simple comprehensive method for functional definition, isolation and characterization of STS for genome mapping**. In: *Human Genome II: The international conference on the status and future of research on the Human Genome*; 1990; San Diego, CA.